SPECIAL ISSUE ON CONTROL AND OPTIMIZATION IN COOPERATIVE NETWORKS

This special issue of the *SIAM Journal on Control and Optimization* is motivated by the emerging disciplines of multi-agent networks, distributed motion coordination, and cooperative control. It is foreseen that, in the near future, large numbers of coordinated devices communicating through ad hoc networks will perform a variety of challenging sensing tasks. The potential advantages of employing arrays of sensors are numerous: certain tasks are difficult, if not impossible, when performed by a single agent; and a group of agents inherently provides robustness to failures of single agents or communication links. Although the individual components of these networked systems are increasingly sophisticated, we lack a fundamental understanding of how to assemble and coordinate the individual physical devices into a coherent whole. As a consequence of this limitation, there exists a strong need for the integration of the sensing, computing and networking aspects of coordinated control of networks.

This special issue gathers recent developments aimed at providing new sets of control and distributed optimization tools to address the research challenges posed by multi-agent networks. This volume presents 16 papers that deal with a wide range of coordination tasks such as consensus, connectivity maintenance, spatial process estimation, formation stabilization, localization, coverage, and target detection. The topics considered include the characterization of convergence speeds, the design of algorithms that tolerate delays, noisy measurements, and packet drops, and the study of the controllability of graph structures and the influence of the interconnection topology in the control design. Collectively, the papers detail deep connections between a wide variety of scientific disciplines, including cooperative, distributed and decentralized control theory, distributed algorithms and systems, communication networks, graph theory, geometric optimization, differential and computational geometry, game and learning theory, and process estimation.

We believe this special issue contains an excellent sample of current research on multi-agent networks and we envision that it will be of great interest to the broad readership of the *SIAM Journal on Control and Optimization.*

A million thanks are due to all the authors who submitted their work to this special issue. We would also like to thank the proficient and timely editorial support provided by Brian Fauth and Mitch Chernoff and the insightful and gracious advice that we received from the editor-in-chief, John Baillieul.

Francesco Bullo
University of California, Santa Barbara

Jorge Cortés
University of California, San Diego

Benedetto Piccoli
Istituto per le Applicazioni del Calcolo "Mauro Picone"
Guest editors

SIAM J. Control Optim.
Vol. 48, No. 1, pp. 1–32

© 2009 Society for Industrial and Applied Mathematics

CONVERGENCE SPEED OF UNSTEADY DISTRIBUTED CONSENSUS: DECAY ESTIMATE ALONG THE SETTLING SPANNING-TREES[*]

DAVID ANGELI[†] AND PIERRE-ALEXANDRE BLIMAN[‡]

Abstract. Results for estimating the convergence rate of nonstationary distributed consensus algorithms are provided, on the basis of qualitative (mainly topological) as well as basic quantitative information (lower-bounds on the matrix entries). The results appear to be tight in a number of instances and are illustrated through simple as well as more sophisticated examples. The main idea is to follow propagation of information along certain spanning-trees which arise in the communication graph.

Key words. distributed consensus, multiagent systems, convergence rate, graph theory

AMS subject classifications. 93C05, 05C50, 05C90, 93C10, 93C55, 93D20, 98R10

DOI. 10.1137/060673527

1. Introduction.

Historically appearing in the areas of communication networks, control theory, and parallel computation, the analytical study of ways for reaching consensus in a population of agents is a problem of broad interest in many fields of science and technology. Questions of this nature arise in peer-to-peer and sensor networks [8, 1], in the maneuvring of groups of vehicles [2, 23, 34], in the study of transmission control protocol (TCP) protocols [5], in the theory of coupled oscillators [16, 25, 21, 31], in neural networks [20], but also in apparently distant fields such as the study and modeling of opinion in social science [24] and of animal flocking [13].

Generally speaking, the goal of such studies is to design or analyze decentralized algorithms through which "agents" (which in the previous examples can be cars or unmanned aerial vehicles, nodes in communication network, sensors, particles, cells, fish, etc.) can update their internal states in order to agree on a common value. In general, the latter shall not be a priori fixed but will be determined as a result of the interactions and of their history.[1] These interactions can be modeled either as unidirectional or bidirectional, corresponding to different extreme situations in which one agent is able to influence another without being affected by the internal state of the receiver (as in hierarchical communication flows) or, conversely, in which influence between agents is always symmetrical.

Of particular interest is the question of estimating how quickly consensus is reached on the basis of qualitative (mainly topological) as well as basic quantitative information (strength of reciprocal influences) on the network.

Originally, the problem of quantifying the convergence rate towards consensus was considered in the context of stationary networks. For Markov chains, for example,

[*] Received by the editors October 27, 2006; accepted for publication (in revised form) April 7, 2008; published electronically February 11, 2009.
http://www.siam.org/journals/sicon/48-1/67352.html

[†] Dipartimento di Sistemi e Informatica, University of Florence, Via di S. Marta 3, 50139 Firenze, Italy (angeli@dsi.unifi.it), and Department of Electrical and Electronic Engineering, Imperial College, London, UK.

[‡] INRIA, Rocquencourt BP105, 78153 Le Chesnay cedex, France (pierre-alexandre.bliman@inria.fr).

[1] Distributed averaging corresponds to the special case where the limit is guaranteed to be the average of the individual values.

this amounts to quantifying the speed at which steady-state probability distribution is achieved, and is therefore directly related to finding an a priori estimate to the second largest eigenvalue of a stochastic matrix. Classical works on this subject are due to Cheeger [11] and Diaconis and Stroock [14]; see also [15] for improved bounds.

Among the classical contributions which instead deal with time-varying interactions, we refer to the work of Cohn [12], where asymptotic convergence is proved, but where the issue of relating topology and guaranteed convergence rates is neglected. Tsitsiklis and coworkers also provided important qualitative contributions to this subject [32, 33, 6], as has Moreau [28]. See also [3] for further nonlinear results. In particular, the role of connectivity of the communication graph in the convergence of consensus has been recognized and finely analyzed.

As noticed in different manners by the preceding authors, arguments based on graph theory are more powerful and seem to catch, in a more natural way, the essence of the problem, rather than computations based on linear algebra techniques (although the study of stochastic matrices offers nowadays, undoubtedly, quite strong results). We are in perfect harmony with the opinion that a view in terms of graphs is central to understanding the agreement issues. However, it appears that some dynamical aspects which have been so far disregarded can be exploited to really gain a tighter understanding of how rapidly consensus can be reached. Our attempt here is to provide a subsequent step towards integration of the temporal aspects of information transit. We are thus led to further elaborate on and exploit tools for description of the connectivity emergence in the communication graphs.

Our purpose in this paper is to provide several criteria to estimate quantitatively the contraction rate of a set of agents towards consensus, in a discrete time framework. Using the language of dynamical systems, the problem here is of estimating the second largest Lyapunov exponent of an infinite product of matrices (see also [7] for links to joint spectral radius). To the best of our knowledge, previous results are centrally based on the existence of a lower-bound of the nonzero entries associated to such matrices, with most of them on the existence of self-loops; see [10] and the surveys in [7, 30] (see, however, the contributions in [32, 33], where the assumption on self-loops is relaxed). Recently, Nedich and Ozdaglar [29] proposed improved bounds under similar assumptions. In contrast, we attempt here to follow more closely the spread of the information over the agent population, along one or more spanning-trees.

Ensuring a lower bound to the matrix entries of the agents already attained by the information flow along the spanning-tree, rather than all the nonzero contributions as classically, permits us to obtain tighter estimates with weaker assumptions. The setting used here applies indifferently to leader-follower or to leaderless networks.

More precisely, our main idea is to examine the birth and rise of spanning-trees in the network. Distinguishing between different subpopulations of agents already touched by spanning-trees and of agents not yet attained, and using lower-bounds on the influence of the former ones on the latter ones, one is able to establish rather precise convergence estimates. Due to the nature of the assumptions, the latter possess some intrinsic robustness with respect to parametric uncertainties.

The paper is organized as follows. The problem is formulated in section 2, and some pertinent concepts are therein introduced. Specifically, several appropriate connectivity notions are defined, among them *sequential connectivity*, which turns out to be central to our developments. The remaining sections are devoted to the statement and demonstration of the main results. Section 3 deals with the problem of contraction estimates when information follows a single spanning-tree (or at least one such spanning-tree with a certain guaranteed strength existing in the underlying graph). This section contains remarks on self-loops and delays; see subsection 3.3. The most original results are in section 4, where it is shown, by means of a fairly general technique, how multiple spanning-trees can be used to derive tight estimates of the contraction rate. Conclusions are reported in section 5.

For better readability, various examples are reported in the text to illustrate the application of the results and to demonstrate the strength of the method. Also, some involved results and proofs have been put into appendices. In particular the main technical tool for carrying out estimates over a finite horizon of the contraction rate of a linear stochastic system is given there.

Notation. In what follows, $\mathbb{N}$ stands for the set of natural integers (including zero), and $\lfloor x \rfloor$ designates the integer value of a real number x. For any set $\mathcal{N}$, we denote by $|\mathcal{N}|$ or card $\mathcal{N}$ its cardinality. Generally, Latin or Greek uppercase letters indicate matrices, and lowercase letters are used to signal scalar numbers and vectors. Graphs and sets are distinguished by calligraphic letters.

In what follows, the (possibly infinite-dimensional) vectors $\mathbf{1}$ and $\mathbf{1}_i$ denote, respectively, a column of 1 and the vector with null components, except 1 in the ith position.

We call an *integer interval* any set obtained as the intersection of a usual interval with the set $\mathbb{N}$. When the context is clear, in particular, when talking about time values, the integer intervals are denoted as the classical ones, for example, $[0, T] \doteq \{t \in \mathbb{N} \ : \ 0 \leq t \leq T\}$.

For p, q positive integers, we denote by I_p and $0_{p\times q}$ the identity and zero matrices, respectively. The transposition of matrices is denoted $^{\intercal}$. By definition, (row) stochastic (resp., substochastic) matrices are square matrices with nonnegative components, whose row sums are equal (resp., at most equal) to 1. Their spectrum is ordered by nonincreasing modulus magnitude, i.e., for M stochastic in $\mathbb{R}^{n\times n}$, $1 = \lambda_1(M) \geq |\lambda_2(M)| \geq \cdots \geq |\lambda_n(M)|$.

Lastly, we introduce the matrix sets $\mathbb{M}^{p,q}$. By definition,

$$\mathbb{M}^{p,q} \doteq \left\{M \in \mathbb{R}^{p\times q} \ : \ M \geq 0 \text{ and } \forall i = 1, \dots, p,\ M_{i,1} + \cdots + M_{i,q} \geq 1\right\} \ . \tag{1}$$

In (1) and throughout the paper, matrix ordering is meant componentwise, i.e., $M \geq 0$ stands for $M_{i,j} \geq 0$ for all i, j.

2. Sequential connectivity and other graph-related notions. We consider the problem of convergence of the consensus algorithm described by the following system:

$$x_k(t+1) = \sum_{l\in\mathcal{N}} \gamma_{k,l}(t) x_l(t),\ k \in \mathcal{N} \ , \tag{2}$$

towards a *common value*; that is, the *global asymptotic stability of the diagonal set* $\{x \ : \ \text{for all } k, l \in \mathcal{N}, x_k = x_l\}$. As usual, (2) may be written in matrix form as

$$x(t+1) = \Gamma(t)x(t), \quad x(t) \doteq \begin{pmatrix} x_1(t) \\ \vdots \end{pmatrix}, \quad \Gamma(t) \doteq (\gamma_{k,l}(t))_{(k,l)\in\mathcal{N}\times\mathcal{N}} \ .$$

We consider scalar systems, although extension to multidimensional systems is possible. The set $\mathcal{N}$ is finite or countable, and the functions x_k map $\mathbb{N}$ to $\mathbb{R}$. We assume throughout that

$$\forall k, l \in \mathcal{N}, \forall t \in \mathbb{N},\ \gamma_{k,l}(t) \geq 0, \text{ and } \forall k \in \mathcal{N},\ \sum_{l\in\mathcal{N}} \gamma_{k,l}(t) = 1 \ . \tag{3}$$

In other words, the matrices $(\gamma_{k,l}(t))_{(k,l)\in\mathcal{N}\times\mathcal{N}}$ are stochastic.

Our goal in the remainder of the paper is to quantify the convergence speed of the set $\{x_k(t) \ : \ k\in\mathcal{N}\}$ when $t\to+\infty$ towards a consensus value. We first introduce vocabulary adequate to measure the latter.

Definition 1 (agent set diameter). *The quantity*

$$\Delta(x(t)) \doteq \sup_{k\in\mathcal{N}} x_k(t) - \inf_{k\in\mathcal{N}} x_k(t)$$

is called the diameter of the agents set at time t.

In what follows, $\Delta(x(t))$ plays the role of a Lyapunov function to study convergence to an agreement. Although the latter depends upon the state, we frequently abbreviate the notation in $\Delta(t)$ if no misinterpretation is possible.

Definition 2 (contraction rate). *We call the* contraction rate *of system* (2) *the number* $\rho\in[0,+\infty]$ *defined as*

$$\rho \doteq \sup_{x(0)} \limsup_{t\to+\infty} \left(\frac{\Delta(t)}{\Delta(0)}\right)^{\frac{1}{t}} .$$

The number ρ is indeed the second largest Lyapunov exponent of the dynamical system (2).

Some notions and definitions necessary to describe pertinent aspects of the communication between the agents are now introduced, based on some elementary tools of algebraic graph theory.

Definition 3 (communication graph). *We denote by* communication graph *(of system* (2)*)* at time t *the directed graph defined by the ordered pairs* $(k,l)\in\mathcal{N}\times\mathcal{N}$ *such that* $\gamma_{k,l}(t)>0$.

In the present context, we use the terms "node" and "agent" interchangeably.

Definition 4 (neighbors). *Given a graph* $\mathcal{A}$ *and a nonempty subset* $\mathcal{L}\subseteq\mathcal{N}$, *the set* Neighbors$(\mathcal{L},\mathcal{A})$ *of* neighbors of $\mathcal{L}$ *is the set of those agents* $k\in\mathcal{N}\setminus\mathcal{L}$ *for which there exists at least one element* $l\in\mathcal{L}$ *such that* $(k,l)\in\mathcal{A}$. *When* $\mathcal{L}$ *is a singleton* $\{l\}$, *the notation* Neighbors$(l,\mathcal{A})$ *is used instead of* Neighbors$(\{l\},\mathcal{A})$.

A key property, namely *weak connectivity*, has been shown to crucially influence the evolution of finite systems of agents linked by time-varying communication graphs (see [26, 28]; also see [9], where the weakly connected sequences are called "repeatedly jointly rooted").

Definition 5 (connectivity and weak connectivity). *A node* $k\in\mathcal{N}$ *is said to be* connected *to a node* $l\in\mathcal{N}$ *on a directed graph* $\mathcal{A}$ *defined on* $\mathcal{N}$ *if there exists a* path *joining* k *to* l *in* $\mathcal{A}$ *and respecting the orientation of the arcs. Given a sequence of directed graphs* $\mathcal{A}(t)$, $t\in\mathbb{N}$, *the node* $k\in\mathcal{N}$ *is said to be* connected *to the node* $l\in\mathcal{N}$ on an integer interval $I\subseteq\mathbb{N}$ *if* k *is connected to* l *for the graph* $\bigcup_{t\in I}\mathcal{A}(t)$.

A graph $\mathcal{A}$ *is called* weakly connected [26] *if there is a node* $k\in\mathcal{N}$ *connected to all other nodes* $l\in\mathcal{N}$. *A sequence of graphs* $\mathcal{A}(t)$, $t\in\mathbb{N}$, *is called* weakly connected across an integer interval $I\subseteq\mathbb{N}$ *if the graph* $\bigcup_{t\in I}\mathcal{A}(t)$ *is weakly connected (that is, if there is a node connected across* I *to all other nodes). A subgraph connecting an agent to all the other ones is called a* spanning-tree.

The fundamental result found by Moreau states that uniform global asymptotic stability of the set of common equilibria is *equivalent* to the existence of an integer $T>0$ such that the sequence of graphs is weakly connected on any interval of length T [26, 28]. Exponential estimates may be obtained too; see the survey part of [7, 30] and [10, 9]. As a matter of fact, there is no specific difficulty in checking the validity of both these results, with the weaker assumption that the graph sequence is weakly connected on every integer intervals $[t_p, t_{p+1}]$, $p \in \mathbb{N}$, where the t_p define a strictly increasing sequence such that $\limsup_{p\to+\infty} t_{p+1} - t_p \leq T$.

In order to obtain more precise estimates of the decay rate towards consensus value, it is reasonable to introduce some minimal time taken by the information to cover the graph—the preceding connectivity notions were not concerned with the ordering of the arcs constituting the tree. We thus introduce some notions useful for quantifying the minimal time for information spread. This spread plays a central part in the contraction rate estimate to be stated later.

Definition 6 (sequential connectivity of finite graph sequences). *A finite sequence of T graphs with common nodes $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_T$ ($T \in \mathbb{N}$) is said to be* sequentially connected *if there exist a node $k \in \mathcal{N}$ and iterations given by*

$$\mathcal{N}_0 = \{k\},$$
$$\mathcal{N}_t \subseteq \mathcal{N}_{t-1} \cup \text{Neighbors}(\mathcal{N}_{t-1}, \mathcal{A}_t) \quad t = 1, \ldots, T,$$

which satisfy $\mathcal{N}_T = \mathcal{N}$.

When we want to emphasize the "root" node, we denote $\mathcal{N}_t$ by $\mathcal{N}_t(k)$, meaning that the iteration departs from node k.

The sets introduced in Definition 6 are crucial to understanding the principle of the method developed in the present paper. For each $t = 1, \ldots, T$, the set $\mathcal{N}_t$ contains agents already in $\mathcal{N}_{t-1}$ and agents having a neighbor in $\mathcal{N}_{t-1}$ at time t; i.e., they are all agents which have been attained at most at time t by the settling of the spanning-tree rooted in k.

We now introduce a derived notion for infinite sequences of graphs.

Definition 7 (T-sequential connectivity). *An infinite sequence of graphs $\mathcal{A}(t)$, $t \in \mathbb{N}$, is said to be* T-sequentially connected *if there exists a strictly increasing integer sequence t_p, $p \in \mathbb{N}$, fulfilling*

$$\limsup_{p\to+\infty} t_{p+1} - t_p \leq T \tag{4}$$

and such that each graph subsequence

$$\mathcal{A}(t_p), \ldots, \mathcal{A}(t_{p+1} - 1)$$

is sequentially connected.

Note that the property is, by definition, monotone with respect to T, viz.,

$$T\text{-sequential connectivity} \Rightarrow (T+1)\text{-sequential connectivity.}$$

Moreover, T-sequential connectivity is invariant with respect to finite time shifts, namely, $\mathcal{A}(t)$ is T-sequentially connected iff for all $q \in \mathbb{N}$, $\mathcal{A}(t+q)$ is again T-sequentially connected. Similarly, T-sequential connectivity is invariant with respect to deletions and/or substitutions of finitely many graphs in a sequence, thus confirming that the property is truly an asymptotic definition.

Notice the proximity of the definitions of sequential connectivity proposed here to the notion of weak connectivity; the central difference is that the former takes into account explicitly the time scheduling of the information transit. The following result links the different connectivity properties defined above and provides mutual bounds between the different connectivity time constants. Its proof, due to its simplicity, is omitted for sake of space and can be found in [4, Proposition 3].

PROPOSITION 8. *Any T-sequentially connected sequence of graphs is weakly connected on the integer intervals $[t_p, t_{p+1}]$, $p \in \mathbb{N}$. Reciprocally, given an increasing sequence t_p fulfilling* (4)*, any sequence of graphs defined on a set of n agents that is weakly connected on the intervals $[t_p, t_{p+1}]$, $p \in \mathbb{N}$, is $(n-1)^2T$-sequentially connected.*

3. Propagation of a unique spanning-tree.

3.1. Estimating the contraction: A key lemma. A first result is now given, describing the elementary mechanism which permits us to quantify a contraction along a *unique* spanning-tree.

LEMMA 9. *Let the finite sequence of communication graphs $\mathcal{A}(0), \dots, \mathcal{A}(T-1)$ of system* (2) *be sequentially connected, and let $\mathcal{N}_0, \dots, \mathcal{N}_T$ be the sets corresponding to the spanning-tree (see Definition* 6*). Assume that, for any $t = 0, \dots, T-1$ and any $k \in \mathcal{N}$,*

$$k \in \mathcal{N}_{t+1} \Rightarrow \sum_{l \in \mathcal{N}_t} \gamma_{k,l}(t) \geq \alpha(t) \tag{5}$$

for a given map $\alpha : [0, T-1] \to [0,1]$. Then

$$\Delta(T) \leq \left(1 - \prod_{t=0}^{T-1} \alpha(t)\right) \Delta(0) \ . \tag{6}$$

Besides sequential connectivity, it is thus assumed in Lemma 9 that, when an agent is in $\mathcal{N}_{t+1}$ (and thus is attained by the spanning-tree at time at most $t+1$), at time t the total weight in the right-hand side of (2) of its neighbors from $\mathcal{N}_t$ (which have been previously attained by the spanning-tree), including possibly itself, is at least $\alpha(t)$ until completion of the tree. This is thus a hypothesis on the relative value of the two “feeding weights” internal and external to the spanning-tree.

Proof of Lemma 9. Lemma 9 is a particular case of a more complex result, Lemma 18, which will be used and demonstrated further. For this reason, we limit the present proof to the essential arguments. For any $k \in \mathcal{N}_{t+1}$ we have

$$x_k(t+1) = \sum_{l=1,\dots,n} \gamma_{k,l}(t) x_l(t) = \sum_{l \in \mathcal{N}_t} \gamma_{k,l}(t) x_l(t) + \sum_{l \in \mathcal{N} \setminus \mathcal{N}_t} \gamma_{k,l}(t) x_l(t) \ ,$$

where by assumption, $\sum_{l \in \mathcal{N}_t} \gamma_{k,l}(t) \geq \alpha(t)$ and $\sum_{l=1,\dots,n} \gamma_{k,l}(t) = 1$. From this, it may be shown that

$$\sup_{k \in \mathcal{N}_{t+1}} x_k(t+1) \leq \alpha(t) \sup_{k \in \mathcal{N}_t} x_k(t) + (1 - \alpha(t)) \sup_{k \in \mathcal{N}} x_k(t) \ ,$$

and an opposite inequality can be shown for the corresponding inf expressions. Denoting

$$\Delta_1(t) \doteq \sup_{k \in \mathcal{N}_t} x_k(t) - \inf_{k \in \mathcal{N}_t} x_k(t) \ ,$$

it turns out that $\Delta_1(0) = 0$ ($\mathcal{N}_0$ is a singleton, the root of the spanning-tree), while $\Delta_1(T) = \Delta(T)$ (when the spanning-tree has run along the entire graph at time T). Thus

$$\Delta(T) = \Delta_1(T) \leq (1 - \alpha(0)\alpha(1)\dots\alpha(T-1))\ \Delta(0),$$

as claimed in the statement. □

Remark 1. Under the hypotheses of Lemma 9, one may show easily that $x(T)$, considered as a function of $x(0)$, verifies

$$\forall l \in \mathcal{N},\ \frac{\partial x_l(T)}{\partial x_k(0)} \geq \prod_{t=0}^{T-1} \alpha(t)\ , \tag{7}$$

where k denotes the index of the root of the spanning-tree. Indeed, one has more generally, for any $t = 0, \dots, T-1$,

$$\forall l \in \mathcal{N}_{t+1},\ \frac{\partial x_l(t+1)}{\partial x_k(0)} \geq \alpha(0)\alpha(1)\dots\alpha(t)\ .$$

From (7), one deduces that, for any $l \in \mathcal{N}$,

$$x_l(T) = \prod_{t=0}^{T-1} \alpha(t)\ x_k(0) + \sum_{l' \in \mathcal{N}} \zeta_{l,l'}(0,T) x_{l'}(0)\ .$$

Here the $\zeta_{l,l'}(0,T)$ are some nonnegative real coefficients, whose explicit value is not needed, i.e., the previous formula just indicates that $x_l(T)$ is a convex combination of the components $x_{l'}(0)$, with a minimal weight on $x_k(0)$. In particular, for any $l \in \mathcal{N}$, the weight of the remaining terms verifies

$$\sum_{l' \in \mathcal{N}} \zeta_{l,l'}(0,T) = 1 - \prod_{t=0}^{T-1} \alpha(t)\ .$$

It is then immediate to establish that

$$\begin{aligned}\left(1 - \prod_{t=0}^{T-1} \alpha(t)\right) \inf_{l \in \mathcal{N}} x_l(0) \leq \inf_{l \in \mathcal{N}} x_l(T) - \prod_{t=0}^{T-1} \alpha(t)\ x_k(0) \\ \leq \sup_{l \in \mathcal{N}} x_l(T) - \prod_{t=0}^{T-1} \alpha(t)\ x_k(0) \leq \left(1 - \prod_{t=0}^{T-1} \alpha(t)\right) \sup_{l \in \mathcal{N}} x_l(0)\ ,\end{aligned}$$

which furnishes an alternative proof of inequality (6). One can thus interpret assumption (5) as ensuring a minimal guaranteed influence of the value of $x_k(0)$ (at the root of the spanning-tree) on every value $x_l(T)$, $l \in \mathcal{N}$. □

3.2. Results on the contraction rate estimate. The main result of section 3 is now presented. A direct consequence of Lemma 9, it provides an estimate of the contraction rate.

THEOREM 10. *Let the sequence of communication graphs of system* (2) *be T-sequentially connected. Accordingly, denote by t_p the corresponding increasing integer sequence of spanning-tree completion (see Definition* 7*); by $\mathcal{N}_{p,t-t_p}$, $t = t_p, t_p + 1, \dots, t_{p+1}$ the sets corresponding to the spanning-tree connecting sequentially the graph subsequences $\mathcal{A}(t_p), \dots, \mathcal{A}(t_{p+1} - 1)$, $p \in \mathbb{N}$ (see Definition* 6*); and by $\mathcal{T}$ the corresponding set*

$$\mathcal{T} \doteq \{(p,t)\ :\ p \in \mathbb{N}, t \in \{t_p, \dots, t_{p+1} - 1\}\}\ . \tag{8}$$

Assume existence of a map $\alpha : \mathcal{T} \to [0,1]$ such that, for any $(p,t) \in \mathcal{T}$, for any $k \in \mathcal{N}$,

$$k \in \mathcal{N}_{p,t-t_p+1} \Rightarrow \sum_{l \in \mathcal{N}_{p,t-t_p}} \gamma_{k,l}(t) \geq \alpha(p,t) \ . \tag{9}$$

Then the contraction rate of system (2) *as defined in Definition* 2 *verifies*

$$\rho \leq \limsup_{p\to+\infty} \prod_{p'=1}^{p} \left(1 - \prod_{t=t_{p'}}^{t_{p'+1}-1} \alpha(p',t) \right)^{1/t_{p+1}} \ . \tag{10}$$

Notice that, with the definition adopted in (8), there is indeed, for each $t \in \mathbb{N}$, a unique $p \in \mathbb{N}$ such that $(p,t) \in \mathcal{T}$.

An important feature is that self-loops ($\gamma_{k,k} > 0$) are not mandatory here, contrary to other previous contributions; see [6, 28, 7]. This assumption is loosened up in [32, 33] for some, but not all, agents. Example 3 below presents an example where this is further weakened. In particular, this feature permits us to model leader/follower evolutions as well as leaderless networks within a unified framework. On this subject, see also subsection 3.3 below.

Similarly, no positive uniform lower-bound on the nonzero coefficients of $\Gamma(t)$ is required; i.e., requirement (9) is sensibly weaker than the usual one in the literature; see [10, 9, 30] and our Example 1.

Proof of Theorem 10. One first states a monotonicity result for the diameter of the agent set along the solutions of (2).

Lemma 11. *For any trajectory of* (2)*, one has, for any $t \in \mathbb{N}$,*

$$\Delta(t+1) \leq \Delta(t) \ .$$

Proof. The proof of Lemma 11 comes from the fact that, the matrices $\Gamma(t)$ being stochastic, the map $t \mapsto \sup_{k\in\mathcal{N}} x_k(t)$ (resp., $t \mapsto \inf_{k\in\mathcal{N}} x_k(t)$) is nonincreasing (resp., nondecreasing). □

One deduces directly from (6) that

$$\Delta(t_{p+1}) \leq \prod_{p'=1}^{p} \left(1 - \prod_{t_{p'}}^{t_{p'+1}-1} \alpha(p',t) \right) \Delta(t_1) \ .$$

Thus,

$$\begin{aligned}\limsup_{p\to+\infty} e^{\frac{1}{t_{p+1}} \ln\left(\frac{\Delta(t_{p+1})}{\Delta(0)}\right)} &= \limsup_{p\to+\infty} e^{\frac{1}{t_{p+1}} \ln\left(\frac{\Delta(t_1)}{\Delta(0)}\right)} e^{\frac{1}{t_{p+1}} \ln\left(\frac{\Delta(t_{p+1})}{\Delta(t_1)}\right)} \\ &\leq \limsup_{p\to+\infty} \prod_{p'=1}^{p} \left(1 - \prod_{t=t_{p'}}^{t_{p'+1}-1} \alpha(p',t) \right)^{1/t_{p+1}} \ .\end{aligned}$$

Clearly,

$$\limsup_{p\to+\infty} e^{\frac{1}{t_{p+1}} \ln\left(\frac{\Delta(t_{p+1})}{\Delta(0)}\right)} \leq \limsup_{t\to+\infty} e^{\frac{1}{t} \ln\left(\frac{\Delta(t)}{\Delta(0)}\right)} \ .$$

Now, from the fact that $\Delta(t)$ is nonincreasing and $t_{p+1}/t_{p+2} \to 1$ as $p \to +\infty$, one gets

$$\limsup_{t\to+\infty} e^{\frac{1}{t}\ln\left(\frac{\Delta(t)}{\Delta(0)}\right)} \leq \limsup_{p\to+\infty} e^{\frac{1}{t_{p+2}}\ln\left(\frac{\Delta(t_{p+1})}{\Delta(0)}\right)} = \limsup_{p\to+\infty} e^{\frac{1}{t_{p+1}}\ln\left(\frac{\Delta(t_{p+1})}{\Delta(0)}\right)}$$

(notice that the logarithmic expressions are not positive, due to the nonincreasingness of Δ along time). The conclusion is then immediate from the definition of ρ given in Definition 2. □

The next result is a specialization of Theorem 10 for constant α.

Corollary 12. *Let the sequence of communication graphs of system* (2) *be T-sequentially connected. Assume the existence of a constant map α in* $[0,1]$ *satisfying* (9). *Then*

$$\rho \leq (1-\alpha^T)^{\frac{1}{T}} \ . \tag{11}$$

The previous results extend similar estimates found previously (see [6, 10, 9, 29]), as α does not have to bound from below the components of the matrices $\Gamma(t)$.

Proof of Corollary 12. Assume without loss of generality $t_{p+1} - t_p \leq T$ for all $p \in \mathbb{N}$. Consequently $\limsup_{p\to+\infty} pT/t_{p+1} \geq 1$. Applying Theorem 10 with constant α yields, for every $p \in \mathbb{N}$,

$$\begin{aligned}\rho \leq \limsup_{p\to+\infty} \prod_{p'=1}^{p} \left(1-\alpha^{t_{p'+1}-t_{p'}}\right)^{1/t_{p+1}} &\leq \limsup_{p\to+\infty} \prod_{p'=1}^{p} \left(\left(1-\alpha^T\right)^{1/T}\right)^{T/t_{p+1}} \\ &= \limsup_{p\to+\infty} \left(\left(1-\alpha^T\right)^{1/T}\right)^{pT/t_{p+1}} \leq (1-\alpha^T)^{1/T},\end{aligned}$$

where one has used the fact that $T \mapsto (1-\alpha^T)$ is increasing on $\mathbb{R}^+$ for any $\alpha \in [0,1]$. Corollary 12 is thus proved. □

Remark 2. A classical topic in linear algebra is the estimate of the second largest eigenvalue (in modulus) of a stochastic matrix for large dimensions. In particular, as n grows, Landau and Odlyzko [22] showed that the rate of convergence is of order $1 - 1/n^3$ (with n being the number of agents) for the equal neighbor time invariant model on an undirected graph; see also results of the same nature in [30]. Our results can be applied to large systems as well. In particular, each given topology induces some kind of relation (typically an inequality) between tree-depth, weight of edges, and number of agents. This inequality can, in principle, be used to derive convergence rate estimates based on the number of agents. □

We now provide several examples of systems with $n = 3$ agents in order to illustrate the two previous results.

Example 1. As a first example, consider the stationary system with $n = 3$ agents given by

$$\Gamma = \Gamma(\varepsilon) \doteq \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 2/3-\varepsilon & \varepsilon \end{pmatrix}$$

for fixed $\varepsilon \in [0, 1/3]$. For $\varepsilon = 1/3$, we obtain the equal neighbor averaging model corresponding to a complete graph [30]. A spectral analysis argument shows that the actual value of the contraction rate ρ is equal to $1/3 - \varepsilon$. Taking into account the fact that the coefficients are greater than or equal to $\min\{1/3, 2/3-\varepsilon, \varepsilon\} = \varepsilon$, methods in [10, 9, 30] yield an upper estimate of ρ equal to $1-\varepsilon^2$, or even $1-\varepsilon \geq 2/3$ (taking into account the fact that the system under study is neighbor shared [10, 9] and adapting [10, Lemma 2 and Theorem 1] to systems whose nonzero coefficients are at least ε). Now, Corollary 12 can be applied. Indeed, the system appears as 1-*sequentially connected*—as the first node participates with nonzero weight to the evolution of all the agents—and one can take $\alpha = 1/3$ as a lower-bound for these weights. This gives an estimate of ρ equal to 2/3, which is better than the results obtained by the other methods. □

Example 2. Consider a system with $n = 3$ agents and dynamics defined by (2) with

(12)
$$\Gamma(t) \doteq \left(\begin{array}{c|cc} \alpha(t)M_1(t) & \begin{matrix}\star & \star\\ \star & \star\end{matrix} \\ \hline \star & \star \ \ \star \end{array}\right) \text{ if } t \in 2\mathbb{N},\ \Gamma(t) \doteq \left(\begin{array}{c|c} \alpha(t)M_2(t) & \begin{matrix}\star\\ \star\\ \star\end{matrix} \end{array}\right) \text{ if } t \in 2\mathbb{N}+1\ .$$

Here, $\alpha : \mathbb{N} \to (0,1]$, $M_1 : \mathbb{N} \to \mathbb{M}^{2,1}$, $M_2 : \mathbb{N} \to \mathbb{M}^{3,2}$ (these sets have been defined in (1)), and the stars stand for any nonnegative scalar numbers rendering the matrix $\Gamma(t)$ stochastic (for this to hold, the row-sums of $M_1(t)$, $M_2(t)$ have to be at most equal to $\alpha(t)^{-1}$).

What is meant here is that for even t, $x_1(t)$ participates with a weight at least $\alpha(t)$ to the value of $x_1(t+1)$ and $x_2(t+1)$, and that for odd t, $x_1(t)$ and $x_2(t)$ participate with a *global* weight at least $\alpha(t)$ to the value of $x_1(t+1)$, $x_2(t+1)$, and $x_3(t+1)$. This is precisely the assumption needed to apply Theorem 10, as detailed now.

Taking the first component x_1 as the root of the spanning-tree, one sees clearly that this system is 2-sequentially connected. Application of Theorem 10 then yields the following estimate:

$$\max_{t'=2t,2t+1}\left\{\max_{i=1,\dots,3} x_i(t') - \min_{i=1,\dots,3} x_i(t')\right\}$$
$$\leq (1-\alpha(0)\alpha(1))\dots(1-\alpha(2t-2)\alpha(2t-1))\left(\max_{i=1,\dots,3} x_i(0) - \min_{i=1,\dots,3} x_i(0)\right)\ ,$$

valid for any $t \in \mathbb{N}$. When α is constant, Corollary 12 applies and leads to

$$\max_{i=1,\dots,3} x_i(t) - \min_{i=1,\dots,3} x_i(t) \leq (1-\alpha^2)^{\lfloor \frac{t}{2} \rfloor}\left(\max_{i=1,\dots,3} x_i(0) - \min_{i=1,\dots,3} x_i(0)\right)\ .$$

The following numerical experiment has been achieved. A set of one thousand couples of stochastic matrices $\Gamma(1)$ and $\Gamma(2)$ are generated randomly (a uniform law on $[0,1]$ is used for each coefficient, and the rows are afterward normalized), and the best estimates for $\alpha(1), \alpha(2)$ fulfilling the conditions above are then computed. The actual contraction rate ρ (which is the square root of the maximal absolute value of the second largest eigenvalues $|\lambda_2(\Gamma(2)\Gamma(1))|$; see [30, Proposition 1]) is then compared to the upper bound $\tilde{\rho}$ deduced from Theorem 10 (that is, $\sqrt{1-\alpha(1)\alpha(2)}$). The corresponding histogram is represented in Figure 1. □

Example 2 shows that, although not tight, the bound may provide reasonable estimates. Notice, however, that the previous comparison test is achieved only with 2-periodic systems (characterized by the second eigenvalue $\lambda_2(\Gamma(2)\Gamma(1))$), although

0–5%	5–10%	10–15%	15–20%	20–25%	25–30%	30–35%
1	21	86	161	212	197	132
35–40%	40–45%	45–50%	50–55%	55–60%	60–65%	65–70%
88	52	30	13	4	2	1

FIG. 1. *Numerical test of Theorem* 10. *Number of occurrences per value of the ratio $\rho/\tilde{\rho}$. See Example* 2 *for details.*

Theorem 10 requires no specific assumption on the general time dependence. An attempt to take into account the occurrence of several spanning-trees is proposed below (section 4).

Example 3. We consider here a simple 2-periodic 3-agent system whose evaluation is not possible by the methods presented by previous works. For $t \in \mathbb{N}$, we let

$$\Gamma(2t) \doteq \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \Gamma(2t+1) \doteq \begin{pmatrix} 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1/2 \\ 1 & 0 & 0 \end{pmatrix}.$$

The matrices $\Gamma(2t+1)$ being deprived of any self-loop, the criteria from [6, 28, 7] cannot be applied. Considering that the system is 2-sequentially connected (with $\mathcal{N}^1_{p,0} = \{1\}$, $\mathcal{N}^1_{p,1} = \{1, 2\}$), using, as in Example 2, Corollary 12 with $\alpha = 1/2$ yields an estimate of the contraction rate as $(1 - 1/4)^{1/2}$, that is, $\sqrt{3}/2 \simeq 0.87$. Indeed, the present example is an instance of Example 2, with $\alpha \equiv \frac{1}{2}$ in (12). On the other hand, using as previously the second eigenvalue argument [30] of the product $\Gamma(2t+1)\Gamma(2t)$, the actual rate is found equal to $\sqrt{5/8} \simeq 0.79$. This value is smaller than $\sqrt{3}/2$, but it is computed under the restrictive hypothesis of periodicity. $\square$

Example 4. As a last illustration of Theorem 10, an elementary time-varying 2-agent system is provided, for which no uniform-in-time lower-bound on the nonzero coefficients of the state matrices exists. This is a situation excluded from the previously published criteria. Let

$$\Gamma(2t) \doteq \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \quad \Gamma(2t+1) \doteq \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ \frac{1}{t} & 1 - \frac{1}{t} & 0 \end{pmatrix}.$$

This is clearly another special instance of Example 2. Theorem 10 applies with $\alpha = 1$ and yields a null contraction rate. Indeed, finite-time convergence does occur, as

$$\Gamma(2t)\Gamma(2t+1) = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}. \qquad \square$$

3.3. Remarks on self-loops and delays. As noticed previously, self-loops are not assumed in the previous results. However, it is a well-known fact that their absence may lead to nonconvergence, as shown, e.g., by the elementary example

$$x_1(t+1) = x_2(t), \quad x_2(t+1) = x_1(t),$$

whose solutions are either constant or oscillating, although the system is undoubtedly sequentially connected.

In fact, the assumptions of the main result of section 3, Theorem 10, force the existence of a self-loop on the root of the spanning-tree at each $t = t_p$, $p \in \mathbb{N}$. This unique self-loop, together with the other hypotheses (which impose, rather, information flux from upstream), turns out to be sufficient to enforce the convergence. After this first step and up to the end of the cycle (that is, for $t_p < t < t_{p+1}$), the root, being in each of the sets $\mathcal{N}_{p,t-t_p+1}$, has to receive a minimal amount of information from within $\mathcal{N}_{p,t-t_p}$, but not specifically via a self-loop. This is, for example, the case in Examples 3 and 4, where for odd t no self-loop occurs for the root of the spanning-tree (agent 1), while the required amount of information is transmitted by agent 2.

A similar remark will hold for the forthcoming results on systems with several spanning-trees given below (see section 4, especially Theorem 16). Notice that the point previously raised is crucial, as it drastically conditions the search for spanning-trees.

On the other hand, it is quite evident that the information transfer between the agents may be subject to delays. This feature does not present specific difficulties a priori, because it can be treated the same way via an augmentation of the state vector. Some past state values are then included in the definition of the diameter and of the contraction rate, which are considered (see Definitions 1 and 2), but this has essentially no consequence on the meaning of this latter quantity.

Generally speaking, delays cannot suppress sequential connectivity, except if they concern the unique mandatory self-loop, located at the root at the initial time instant (see above). On the other hand, they may change the values of the weights $\alpha(t)$ and thus modify the decay estimates. Also, it is rather likely that the decay rate estimates are nondecreasing with respect to any delay.

A more precise study of the quantitative influence of the delays on the convergence speed could be tackled by similar tools, but this feature is beyond the scope of the present paper. For simplicity, we limit ourselves to a simple example, for which analytical results are easily computed.

Example 5. Consider the four following systems:

$$x_1(t+1) = \frac{1}{2}x_1(t) + \frac{1}{2}x_2(t), \quad x_2(t+1) = \frac{1}{2}x_1(t) + \frac{1}{2}x_2(t)\ ; \tag{13a}$$

$$x_1(t+1) = \frac{1}{2}x_1(t) + \frac{1}{2}x_2(t), \quad x_2(t+1) = \frac{1}{2}x_1(t-1) + \frac{1}{2}x_2(t-1)\ ; \tag{13b}$$

(13c)
$$x_1(t+1) = \frac{1}{4}x_1(t) + \frac{1}{4}x_1(t-1) + \frac{1}{2}x_2(t), \quad x_2(t+1) = \frac{1}{2}x_1(t-1) + \frac{1}{2}x_2(t)\ ;$$

$$x_1(t+1) = \frac{1}{2}x_1(t-1) + \frac{1}{2}x_2(t), \quad x_2(t+1) = \frac{1}{2}x_1(t) + \frac{1}{2}x_2(t-1)\ . \tag{13d}$$

The delay-free system (13a) is 1-sequentially connected, and the analysis conducted above yields the estimate $\tilde{\rho} = \frac{1}{2} \geq \rho = 0$.

The three remaining systems possess delayed terms $x_1(t-1)$ and $x_2(t-1)$. Introducing $x_3(t) \doteq x_1(t-1)$, $x_4(t) \doteq x_2(t-1)$, they can be written as $x(t+1) = \Gamma x(t)$, where $x \doteq (x_1, x_2, x_3, x_4)^{\intercal}$ and

$$\Gamma \doteq \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \quad \Gamma \doteq \begin{pmatrix} \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \quad \Gamma \doteq \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix},$$

respectively, for (13b), (13c), and (13d). System (13b) is 3-sequentially connected with $\alpha(0) = \alpha(1) = \alpha(2) = \frac{1}{2}$, so that $\tilde{\rho} = \frac{\sqrt[3]{7}}{2} \simeq 0.96 \geq \rho = \frac{1}{2}$. System (13c) is 2-sequentially connected with $\alpha(0) = \frac{1}{4}$ and $\alpha(1) = \frac{1}{2}$, and $\tilde{\rho} = \frac{\sqrt{7}}{2\sqrt{2}} \simeq 0.94 \geq \rho = \frac{1}{2}$. Lastly, system (13d) is not sequentially connected, due to the absence of a self-loop (zero diagonal). This is corroborated by the fact that $\rho = 1$, so convergence does not occur.

As can be seen, the index of sequential connectivity is not systematically the sum of this index for the delay-free case (1 here) and the sum of the values of the delays.

4. Communication graphs spanned by several spanning-trees.

4.1. Sequential connectivity with several spanning-trees. When several spanning-trees emerge in the communication graph (either simultaneously or successively), the previous analysis may happen to be conservative. We now face the issue of how to tackle this feature.

An extension of the notion of sequential connectivity introduced in section 2 is first constructed, analogously to Definitions 6 and 7.

Definition 13 (sequential connectivity of finite graph sequences by multiple spanning-trees). *A finite sequence of T graphs with common nodes $\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_T$ is said to be* sequentially connected by m spanning-trees *($m \in \mathbb{N}$) if there exist nodes $k^1, k^2, \dots, k^m \in \mathcal{N}$ and iterations given by*

$$\mathcal{N}_0^j = \{k^j\},$$
$$\mathcal{N}_t^j \subseteq \mathcal{N}_{t-1}^j \cup \text{Neighbors}(\mathcal{N}_{t-1}^j, \mathcal{A}_t), \quad t = 1, \dots, T,$$

which satisfy $\mathcal{N}_T^j = \mathcal{N}$ for all $j \in \mathcal{I} \doteq \{1, 2, \dots, m\}$.

Similarly we define the property for infinite graph sequences as follows.

Definition 14 (T-sequential connectivity by multiple spanning-trees). *An infinite sequence of graphs $\mathcal{A}(t)$, $t \in \mathbb{N}$, is said to be T-*sequentially connected by m spanning-trees *if there exists a strictly increasing integer sequence t_p, $p \in \mathbb{N}$, fulfilling* (4) *and such that for all $p \in \mathbb{N}$, each graph subsequence*

$$\mathcal{A}(t_p), \dots, \mathcal{A}(t_{p+1} - 1)$$

is sequentially connected by m spanning-trees.

The following result extends Lemma 9. As the latter, it is directly deduced from Lemma 18, so a detailed proof is omitted.

Lemma 15. *Let the finite sequence of communication graphs $\mathcal{A}(0), \dots, \mathcal{A}(T-1)$ of system* (2) *be sequentially connected by m spanning-trees, and let $\mathcal{N}_0^j, \dots, \mathcal{N}_T^j$ be the sets corresponding to the jth spanning-tree (see Definition* 13*). For each $k \in \mathcal{N}$, consider sets $\mathcal{N}_t^{k,j} \subseteq \mathcal{N}_t^j$ such that $j \neq j' \Rightarrow \mathcal{N}_t^{k,j} \cap \mathcal{N}_t^{k,j'} = \emptyset$. Assume the existence of maps $\alpha_{i,j} : [0, T-1] \to [0, 1]$ such that, for any $t = 0, \dots, T-1$, any $i, j \in \mathcal{I}$, and any $k \in \mathcal{N}$,*

$$k \in \mathcal{N}_{t+1}^i \Rightarrow \sum_{l \in \mathcal{N}_t^{k,j}} \gamma_{k,l}(t) \geq \alpha_{i,j}(t) \tag{14}$$

for given maps $\alpha_{i,j} : [0, T-1] \to [0, 1]$. Then $A(t) \doteq (\alpha_{i,j}(t))_{(i,j) \in \mathcal{I} \times \mathcal{I}}$ is a substochastic matrix in $\mathbb{R}^{m \times m}$ and

$$\Delta(T) \leq \left(1 - \max_{j \in \mathcal{I}} \mathbf{1}_j^{\mathsf{T}} A(T-1) \dots A(0) \mathbf{1}\right) \Delta(0) \ , \tag{15}$$

where $\mathbf{1}$ and $\mathbf{1}_j$ are as defined in the notation section.

The sets $\mathcal{N}_t^j$ generalize the notion introduced in section 3; i.e., here, each set $\mathcal{N}_t^j$ is constituted by agents attained at most at time t by the jth spanning-tree. Assumption (14) fixes a lower-bound $\alpha_{i,j}$ to the total weight applied by each agent in $\mathcal{N}_{t+1}^i$ to agents in $\mathcal{N}_t^j$.

When an agent is a member, for given t, of sets $\mathcal{N}_t^j$ for more than one value of j, it is necessary to decide, in the construction of contraction estimates relative to the x_k update equation, to which of them its influence is attributed; this choice could actually vary according to the considered update equation. This is the reason why subsets $\mathcal{N}_t^{k,j}$ disjoint for different js are introduced.

Remark 3. When in the statement of Lemma 15 the sequential spanning-trees corresponding to two distinct values of i, i' are identical, then the scalar quantities $\mathbf{1}_i^{\mathsf{T}} A(T-1)\dots A(0)\mathbf{1}$ and $\mathbf{1}_{i'}^{\mathsf{T}} A(T-1)\dots A(0)\mathbf{1}$ are equal—at least if the $\alpha_{i,j}(t)$ are chosen to be identical for all $j \in \mathcal{I}$.

On the other hand, different choices in the attribution of arcs to one or another of the m developing spanning-trees (that is, on the definition of the sets $\mathcal{N}_t^{k,j}$) may lead to different choices for these coefficients, and consequently to different estimates. In this respect, adding virtual sequential spanning-trees may allow us to improve the convergence speed estimate; see Example 6 below. □

Remark 4. As in Remark 1 for Lemma 9, one shows easily that, under the hypotheses of Lemma 15, one has similarly, for any $j \in \mathcal{I}$,

$$\forall l \in \mathcal{N},\ \frac{\partial x_l(T)}{\partial x_k(0)} \geq \mathbf{1}_j^{\mathsf{T}} A(T-1)\dots A(0)\mathbf{1}_k\ ,$$

where the index k represents any index of the roots of the m spanning-trees. Indeed, for any $t = 0, \dots, T-1$ and any $j \in \mathcal{I}$,

$$\forall l \in \mathcal{N}_{t+1}^j,\ \frac{\partial x_l(t+1)}{\partial x_k(0)} \geq \mathbf{1}_j^{\mathsf{T}} A(T-1)\dots A(0)\mathbf{1}_k\ ,$$

which gives the previous estimate when $t = T-1$, as $\mathcal{N}_T^j = \mathcal{N}$ for all $j \in \mathcal{I}$. □

4.2. Results on contraction rate estimate. We now come to the key result of section 4, which is also the most powerful result of the paper.

Theorem 16. *Let the sequence of communication graphs of system* (2) *be T-sequentially connected by m spanning-trees. Accordingly, denote by t_p the corresponding increasing integer sequence of spanning-tree completion (see Definition* 14*); by $\mathcal{N}_{p,t-t_p}^j$, $t = t_p, \dots, t_{p+1}$, the sets corresponding to the m spanning-trees connecting sequentially the graph subsequences $\mathcal{A}(t_p), \dots, \mathcal{A}(t_{p+1}-1)$, $p \in \mathbb{N}$ (see Definition* 13*); and by $\mathcal{T}$ the set defined in* (8)*. Furthermore, for each $k \in \mathcal{N}$, consider sets $\mathcal{N}_{p,t-t_p}^{k,j} \subseteq \mathcal{N}_{p,t-t_p}^j$ such that $j \neq j' \Rightarrow \mathcal{N}_{p,t-t_p}^{k,j} \cap \mathcal{N}_{p,t-t_p}^{k,j'} = \emptyset$, and assume the existence of maps $\alpha_{i,j} : \mathcal{T} \to [0,1]$, $(i,j) \in \mathcal{I} \times \mathcal{I}$, such that, for any $(p,t) \in \mathcal{T}$, any $i, j \in \mathcal{I}$, and any $k \in \mathcal{N}$,*

$$k \in \mathcal{N}_{p,t-t_p+1}^i \Rightarrow \sum_{l \in \mathcal{N}_{p,t-t_p}^{k,j}} \gamma_{k,l}(t) \geq \alpha_{i,j}(p,t)\ . \tag{16}$$

Then $A(p,t) \doteq (\alpha_{i,j}(p,t))_{(i,j)\in\mathcal{I}\times\mathcal{I}}$ is a substochastic matrix in $\mathbb{R}^{m\times m}$, and the contraction rate of system (2) *as defined in Definition* 2 *verifies*

$$\rho \leq \limsup_{p\to+\infty} \prod_{p'=1}^{p} \left(1 - \max_{j\in\mathcal{I}} \mathbf{1}_j^{\mathsf{T}} A(p', t_{p'+1}-1)\dots A(p', t_{p'})\mathbf{1}\right)^{1/t_{p+1}}\ . \tag{17}$$

Proof of Theorem 16. Due to the fact that the sets $\mathcal{N}_{k,j}(t)$ are pairwise disjoint for different values of j, one has, for any $i \in \mathcal{I}$, any $t \in \mathbb{N}$,

$$\sum_{j\in\mathcal{I}} \alpha_{i,j}(p, t-t_p) \le \sum_{j\in\mathcal{I}} \sum_{l\in\mathcal{N}_{k,j}(t)} \gamma_{k,l}(t) \le \sum_{l\in\mathcal{N}} \gamma_{k,l}(t) = 1$$

for any $k \in \mathcal{N}^i_{p,t-t_p+1}$. This proves the first part of the statement.

As in the proof of Theorem 10 above, it suffices essentially to establish (15) when t is a multiple of T. Applying Lemma 18 on the integer interval $[t_p, t_{p+1}]$ with $\mathcal{M}_i(t) = \mathcal{N}^i_{p,t-t_p}$, $\mathcal{M}_{k,j}(t) = \mathcal{N}^{k;j}_{p,t-t_p}$, $c_{i,j}(t) = \alpha_{i,j}(p,t)$ yields

$$\begin{aligned}&\Delta_i(t_{p+1})\\ &\quad \le \mathbf{1}_i^\intercal\Big(A(p,t_{p+1}-1)\dots A(p,t_p)\Delta_{\mathcal{N}}(t_p) + \Big(\mathbf{1} - A(p,t_{p+1}-1)\dots A(p,t_p)\mathbf{1}\Big)\Delta(t_p)\Big)\ .\end{aligned}$$

Here, the definition of $\Delta_{\mathcal{N}}$ depends upon p and is as follows:

$$\Delta_{\mathcal{N}}(t) \doteq \begin{pmatrix}\Delta_1(t)\\ \vdots \\ \Delta_m(t)\end{pmatrix},\quad \Delta_i(t) \doteq \sup_{k\in\mathcal{N}^i_{p,t-t_p}} x_k(t) - \inf_{k\in\mathcal{N}^i_{p,t-t_p}} x_k(t),\quad t = t_p,\dots,t_{p+1}\ .$$

By assumption, the existence of the m spanning-trees means that

$$\Delta_i(t_p) = 0 \text{ and } \Delta_i(t_{p+1}) = \Delta(t_{p+1}),\ i \in \mathcal{I}\ .$$

One thus deduces that, for all $i \in \mathcal{I}$,

$$\Delta(t_{p+1}) = \Delta_i(t_{p+1}) \le \mathbf{1}_i^\intercal\left(\mathbf{1} - A(p,t_{p+1}-1)\dots A(p,t_p)\mathbf{1}\right)\Delta(t_p)\ .$$

Thus,

$$\Delta(t_{p+1}) \le (1 - \mathbf{1}_i^\intercal A(p,t_{p+1}-1)\dots A(p,t_p)\mathbf{1})\,\Delta(t_p)\ .$$

The proof is then achieved as it was for Theorem 10. □

Example 6. We come back to the analysis of Example 1, now with the help of Theorem 16. One may distinguish three spanning-trees occurring on each time interval of unit length (in other words, the system is 1-sequentially connected by three spanning-trees), with the root at each of the agents. With this point of view, $\mathcal{I} = \mathcal{N} = \{1,2,3\}$ and $t_p = p$. With the notation of Theorem 16, one may put

$$\mathcal{N}^j_{p,0} = \{j\},\ \mathcal{N}^j_{p,1} = \mathcal{N} = \{1,2,3\},\ \text{for } j = 1,2,3\ .$$

This is the simplest case, where the sets $\mathcal{N}^j_{p,0}$ are pairwise disjoint, so one takes

$$\mathcal{N}^{k,j}_{p,0} \doteq \mathcal{N}^j_{p,0},\ j = 1,2,3\ .$$

We now form the functions $\alpha_{i,j}$, as defined in the statement of Theorem 16, and the corresponding matrix A. By definition, one should have, for any $i, j \in \{1,2,3\}$ (see (16)),

$$\forall k \in \{1,2,3\},\quad \forall p \in \mathbb{N},\quad \gamma_{k,j} \ge \alpha_{i,j}(p,p)\ ,$$

where $\Gamma = (\gamma_{i,j})_{(i,j)\in\mathcal{I}\times\mathcal{I}}$ is as given in Example 1 above. One thus takes

$$\alpha_{i,j}(p,p) \doteq \min_{k=1,2,3} \gamma_{k,j}, \quad i = 1,2,3 ,$$

that is, $\alpha_{i,1} = 1/3$, $\alpha_{i,2} = \min\{1/3, 2/3-\varepsilon\}$, $\alpha_{i,3} = \min\{1/3, \varepsilon\}$, or again

$$A \doteq \begin{pmatrix} 1/3 & \min\{1/3, 2/3-\varepsilon\} & \min\{1/3,\varepsilon\} \\ 1/3 & \min\{1/3, 2/3-\varepsilon\} & \min\{1/3,\varepsilon\} \\ 1/3 & \min\{1/3, 2/3-\varepsilon\} & \min\{1/3,\varepsilon\} \end{pmatrix} .$$

Applying formula (15) then leads to an estimate of the actual contraction rate equal to

$$1 - \Big(1/3 + \min\{1/3, 2/3-\varepsilon\} + \min\{1/3,\varepsilon\}\Big) = 1/3 - \varepsilon .$$

In this example, the method ensuing from Theorem 16 thus generates the exact value of the contraction rate ρ.

Considering now only the two first spanning-trees (with $\mathcal{N}^j_{p,0} = \{j\}$, $\mathcal{N}^j_{p,1} = \mathcal{N} = \{1,2,3\}$ for all $p \in \mathbb{N}$ and all $j \in \{1,2\}$; then $\alpha_{i,1} = 1/3$, $\alpha_{i,2} = \min\{1/3, 2/3-\varepsilon\} = 1/3$, $i = 1,2$) gives a worse estimate, namely $1/3$. Similarly, considering the first and third, or the second and third, spanning-trees yields $2/3-\varepsilon$. These estimates are different and are tighter than $2/3$, the value obtained in Example 1 when considering a unique spanning-tree, but are not optimal. □

We refine further the analysis of systems spanned by several spanning-trees and examine, in subsections 4.3 and 4.4, respectively, the cases of spanning-trees propagating consecutively and simultaneously.

4.3. Application to systems with successive spanning-trees. We now consider the case where several emerging spanning-trees have a common root and possess a certain order property. We mean by this that the dates at which each spanning-tree reaches an agent are interlaced *independently* from the agent. In other words, the "wavefronts" corresponding to each spanning-tree spread in a concentrical manner. Up to renaming, one may label 1 the first spanning-tree, 2 the next one, and so on, and the order property simply reads as follows (reasoning on each interval $[t_p, t_{p+1}]$, we omit the index p):

$$\forall j,j' \in \mathcal{I}, \forall t = 0,\dots,T-1, \quad j \le j' \Rightarrow \mathcal{N}^{j'}_t \subseteq \mathcal{N}^j_t ,$$

and thus, by construction, the following inequalities hold for any $t \in [t_p, t_{p+1}]$:

$$\Delta_m(t) \le \cdots \le \Delta_j(t) \le \cdots \le \Delta_1(t) .$$

It is thus systematically more fruitful to attribute any contribution in the right-hand side of (2) to the set $\mathcal{N}_i$ with largest index i to which it belongs—because the corresponding estimate is tighter. In particular, it is beneficial to choose $\alpha_{i,j} \equiv 0$ for $i < j$, thus leading to lower-triangular matrices A in Theorem 16.

We provide now an illustration of this configuration.

Example 7. For a fixed scalar $\gamma \in [0,1]$, consider the time-invariant system of n agents described by

$$x_1(t+1) = x_1(t), \quad x_i(t+1) = \gamma x_{i-1}(t) + (1-\gamma)x_i(t),\ i = 2,\dots,n .$$

$$A(0)=\begin{pmatrix}\gamma & 0 & \dots & 0\\ 0 & 1 & \dots & 0\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \dots & 1\end{pmatrix},\ A(1)=\begin{pmatrix}\gamma & 0 & \dots & 0\\ 1-\gamma & \gamma & \dots & 0\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \dots & 1\end{pmatrix},\dots,$$

$$A(n-2)=\begin{pmatrix}\gamma & 0 & \dots & 0\\ 1-\gamma & \gamma & \dots & 0\\ 0 & 1-\gamma & \dots & 0\\ \vdots & & & \vdots\\ 0 & 0 & \dots & 1\end{pmatrix},\ A(n-1)=\begin{pmatrix}1 & 0 & \dots & 0\\ 1-\gamma & \gamma & \dots & 0\\ 0 & 1-\gamma & \dots & 0\\ \vdots & & & \vdots\\ 0 & 0 & \dots & 1\end{pmatrix},$$

$$A(n)=\begin{pmatrix}1 & 0 & \dots & 0\\ 0 & 1 & \dots & 0\\ 0 & 1-\gamma & \dots & 0\\ \vdots & & & \vdots\\ 0 & 0 & \dots & 1\end{pmatrix},\dots,\ A(n+q-3)=\begin{pmatrix}1 & 0 & \dots & & \\ 0 & 1 & \dots & & \\ \vdots & & & & \\ 0 & \dots & & 1 & 0\\ 0 & \dots & & 1-\gamma & \gamma\end{pmatrix}.$$

FIG. 2. *Matrices A obtained in Example* 7 *(case $n < q$).*

The corresponding matrix Γ is lower-triangular and admits, apart from 1, a unique eigenvalue, namely $1-\gamma$, with degree $n-1$. The actual value of the contraction rate is thus $\rho = 1-\gamma$.

For any positive integer q, one may consider that the communication graph is spanned by q distinct spanning-trees, departing from agent 1 at time 0, then 1, 2, and so on, up to $q-1$, and attaining agent n at time $n-1$, n, up to $n+q-2$. The duration of this process is thus $T \doteq n+q-2$, and the system may be seen as "T-sequentially connected by q (distinct) spanning-trees." Consistent with the previous notations, we let $t_p = pT$ and consider the sets $\mathcal{N}^i_{p,t}$, $i \in \mathcal{I} \doteq \{1,\dots,q\}$, defined by

$$\mathcal{N}^1_{p,t} = \begin{cases}\{1\} & \text{for } t \le 0,\\ \{1,\dots,t+1\} & \text{for } t = 0,\dots,n-1,\\ \{1,\dots,n\} = \mathcal{N} & \text{for } t = n-1,\dots,n+q-2\end{cases}$$

and

$$\mathcal{N}^{i+1}_{p,t+1} = \mathcal{N}^i_{p,t} \text{ for } i = 2,\dots,q-1\ .$$

Following the progression of each spanning-tree, one shows that one may take for A (in $\mathbb{R}^{q\times q}$) the formulas depicted in Figure 2 (see also Appendix B for details of the proof of Lemma 17).

Let us explain these formulas. From $t=0$ to $t=n-1$, the first spanning-tree spreads from agent 1 to agent n; for the elements of the latter, the right-hand side of the state equation is composed of elements already touched by the information flow (with coefficient γ) and some newly touched element, which consequently does not contribute to the right-hand side. We may therefore choose $\mathcal{N}^{1,1}_{p,t} = \mathcal{N}^1_{p,t}$ and $\mathcal{N}^{1,j}_{p,t} = \emptyset$ for $j = 2\dots q$. This gives rise to the identities $\alpha_{1,1}(t) = \gamma$ and $\alpha_{1,j}(t) = 0$ for $j \in \mathcal{I}\setminus\{1\}$ for $t = 0,\dots,n-2$. At time $t = n-1$, one has $\mathcal{N}^1_{p,t} = \mathcal{N}$ and the expansion of this set is completed, so *all* the terms in the right-hand side come from inside $\mathcal{N}^1_{p,t}$. Thus, again by letting $\mathcal{N}^{1,1}_{p,t} = \mathcal{N}$ and $\mathcal{N}^{1,j}_{p,t} = \emptyset$ for $t = n-1, \dots, n+q-3$ and $j = 2 \dots q$, one has $\alpha_{1,1}(t) = 1$ and $\alpha_{1,j}(t) = 0$.

The second spanning-tree departs from the root at $t = 1$, therefore letting $\mathcal{N}^{2,2}_{p,0} = \mathcal{N}^2_{p,0}$ and $\mathcal{N}^{2,j}_{p,0} = \emptyset$ yields $\alpha_{2,2}(t) = 1$ and $\alpha_{2,j}(t) = 0$ for $j \in \mathcal{I} \setminus \{2\}$ for $t = 0$. Then at $t = 1$, $\mathcal{N}^2_{p,t} = \{1\}$ and $\mathcal{N}^2_{p,t+1} = \{1,2\} = \mathcal{N}^2_{p,t} \cup \left(\mathcal{N}^1_{p,t} \setminus \mathcal{N}^2_{p,t}\right)$. More precisely, the corresponding right-hand side comprises two terms as before: a contribution, with coefficient γ, due to agents already attained by the second spanning-tree, plus a term, with coefficient $1-\gamma$, due to a term coming from an agent not yet touched by the second tree *but already touched by the first one.* We let $\mathcal{N}^{2,1}_{p,1} = \{2\}$, $\mathcal{N}^{2,2}_{p,1} = \{1\}$, $\mathcal{N}^{2,j}_{p,1} = \emptyset$ for $j = 3 \dots q$; this explains that for $t = 1$ one has $\alpha_{2,1}(t) = 1-\gamma$, $\alpha_{2,2}(t) = \gamma$, and $\alpha_{2,j}(t) = 0$ for $j \in \mathcal{I} \setminus \{1,2\}$. Similarly, we define $\mathcal{N}^{2,1}_{p,t} = \{t+1\}$, $\mathcal{N}^{2,2}_{p,t} = \{1 \dots t\}$ and $\mathcal{N}^{2,j}_{p,t} = \emptyset$ for $t = 2 \dots n$, where the second spanning-tree in turn is completed. Again one obtains $\alpha_{2,1}(t) = 1-\gamma$, $\alpha_{2,2}(t) = \gamma$, and $\alpha_{2,j}(t) = 0$ for $j \in \mathcal{I} \setminus \{1,2\}$. Then for subsequent t's, we let $\mathcal{N}^{2,2}_{p,t} = \mathcal{N}$ and $\mathcal{N}^{2,j} = \emptyset$ for $j \neq 2$ so that indeed, $\alpha_{2,2}(t) = 1$ and $\alpha_{2,j}(t) = 0$ for $j \neq 2$.

Lastly, the other spanning-trees appear one by one and share with their predecessor the same relation that the second one shared with the first one. This explains the formulas given, until completion of the qth one, at time $t = T$. The analysis conducted above leads overall to the matrices shown in Figure 2, which corresponds to the case $n < q$ (the first spanning-tree is completed at $t = n$, before the departure of the qth spanning-tree, at $t = q$). The case $n \geq q$ is similar.

For the case of $n = 3$ agents, formula (15) in Theorem 16 then yields the following estimates, denoted $\tilde{\rho}_q$:

- For $q = 1$ (corresponding to the method of Theorem 10),

$$A(0) = A(1) = \gamma \ ,$$

so $\tilde{\rho}_1 = \sqrt{1 - A(1)A(0)} = \sqrt{1-\gamma^2}$.

- For $q = 2$,

$$A(0) = \begin{pmatrix} \gamma & 0 \\ 0 & 1 \end{pmatrix},\ A(1) = \begin{pmatrix} \gamma & 0 \\ 1-\gamma & \gamma \end{pmatrix},\ A(2) = \begin{pmatrix} 1 & 0 \\ 1-\gamma & \gamma \end{pmatrix} \ ,$$

and

$$\begin{aligned} \tilde{\rho}_2 &= \left(1 - \max_{i=1,2} \mathbf{1}_i^{\mathsf{T}} A(2)A(1)A(0)\mathbf{1}\right)^{1/3} \\ &= \left(1 - \max\{\gamma^2; 3\gamma^2 - 2\gamma^3\}\right)^{1/3} = \left(1 - 3\gamma^2 + 2\gamma^3\right)^{1/3} \ . \end{aligned}$$

- For $q = 3$,

$$A(0) = \begin{pmatrix} \gamma & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},\ A(1) = \begin{pmatrix} \gamma & 0 & 0 \\ 1-\gamma & \gamma & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$A(2) = \begin{pmatrix} 1 & 0 & 0 \\ 1-\gamma & \gamma & 0 \\ 0 & 1-\gamma & \gamma \end{pmatrix},\ A(3) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1-\gamma & \gamma \end{pmatrix} \ ,$$

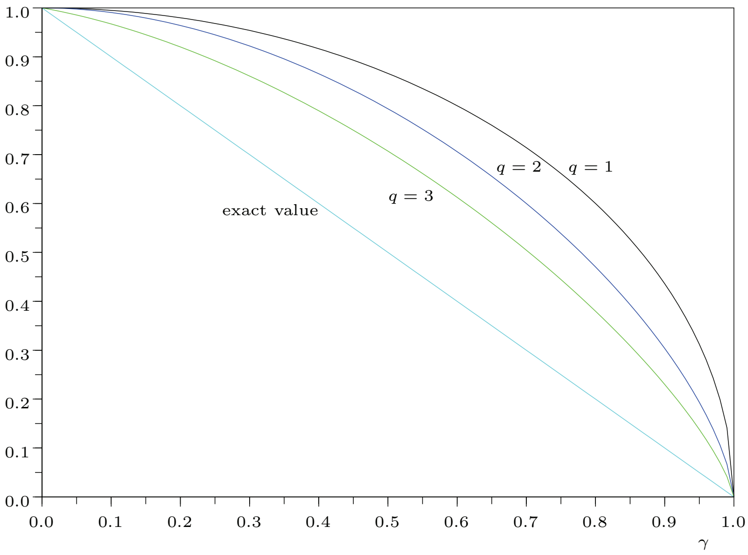


FIG. 3. *Approximations of the contraction rate as functions of γ for different uses of Theorem 16. See Example 7 for details.*

whence

$$\begin{aligned}\tilde{\rho}_3 &= \left(1 - \max_{i=1,2,3} \mathbf{1}_i^{\mathsf{T}} A(3)A(2)A(1)A(0)\mathbf{1}\right)^{1/4} \\ &= \left(1 - \max\{\gamma^2; 3\gamma^2 - 2\gamma^3; 3\gamma^4 - 8\gamma^3 + 6\gamma^2\}\right)^{1/4} = \left(1 - 3\gamma^4 + 8\gamma^3 - 6\gamma^2\right)^{1/4} .\end{aligned}$$

The values obtained approximate the exact value $1 - \gamma$ with increasing precision, as seen in Figure 3. These successive improvements are, of course, the consequence of richer and richer analysis, including more and more settling spanning-trees.

The question of the limiting behavior when q goes to infinity is, of course, intriguing; i.e., is the exact value found asymptotically? It turns out that the answer is positive, as stated now in the general case of a system with n agents.

LEMMA 17. *The value of $\tilde{\rho}_q$ is given by the following formula:*

$$\tilde{\rho}_q^{n+q-2} = 1 - \gamma^{n-1} \sum_{i=0}^{q-1} (1-\gamma)^i \binom{n+i-2}{n-2} = \frac{1}{(n-2)!}\gamma^{n-1} \frac{d^{n-2}}{d\delta^{n-2}} \left[\frac{\delta^{n+q-2}}{1-\delta}\right]\Bigg|_{\delta=1-\gamma} .$$

Consequently, $\tilde{\rho}_q$ tends towards $\rho = 1 - \gamma$ when $q \to +\infty$, and more precisely,

$$\tilde{\rho}_q = \rho + (n-2)(1-\gamma)\frac{\ln q}{q} + o\left(\frac{\ln q}{q}\right) .$$

A proof of Lemma 17 is presented in Appendix B. The calculations have been checked independently by the authors, using a symbolic computation tool.

Although presently limited to special class of examples, Lemma 17 is rather promising; i.e., it establishes that tight estimates may be accessed when employing a large number of settling spanning-trees in the analysis. Extensions are in progress to cover more general cases. □

4.4. Application to systems with concomitant spanning-trees. The examples previously shown drastically exploit the fact that the different spanning-trees occur one after another. We show here that, otherwise, the techniques of Theorem 16 may provide deceivingly weak results.

Example 8. To illustrate this, we consider a system with $n = 6$ agents, T-sequentially connected for $T = 5$. For fixed $\gamma \in (0, 1/2)$, the latter is defined by taking stochastic matrices such that

$$\Gamma(pT) \geq \begin{pmatrix} \gamma & 0 & 0 & 0 & 0 & 0 \\ \gamma & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \; \Gamma(pT+1) \geq \begin{pmatrix} \gamma & 0 & 0 & 0 & 0 & 0 \\ 0 & \gamma & 0 & 0 & 0 & 0 \\ 0 & \gamma & 0 & 0 & 0 & 0 \\ 0 & \gamma & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\Gamma(pT+2) \geq \begin{pmatrix} \gamma & 0 & 0 & 0 & 0 & 0 \\ 0 & \gamma & 0 & 0 & 0 & 0 \\ 0 & 0 & \gamma & 0 & 0 & 0 \\ 0 & 0 & 0 & \gamma & 0 & 0 \\ 0 & 0 & \gamma & 0 & 0 & 0 \\ 0 & 0 & 0 & \gamma & 0 & 0 \end{pmatrix}, \; \Gamma(pT+3) \geq \begin{pmatrix} \gamma & 0 & 0 & 0 & 0 & 0 \\ 0 & \gamma & 0 & 0 & 0 & 0 \\ 0 & 0 & \gamma & 0 & 0 & 0 \\ 0 & 0 & 0 & \gamma & 0 & 0 \\ 0 & 0 & 0 & 0 & \gamma & \gamma \\ 0 & 0 & 0 & 0 & \gamma & \gamma \end{pmatrix},$$

$$\Gamma(pT+4) \geq \begin{pmatrix} \gamma & 0 & 0 & 0 & 0 & 0 \\ 0 & \gamma & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \gamma & 0 \\ 0 & 0 & 0 & 0 & 0 & \gamma \\ 0 & 0 & 0 & 0 & \gamma & 0 \\ 0 & 0 & 0 & 0 & 0 & \gamma \end{pmatrix}$$

for all $p \in \mathbb{N}$ (thus $t_p = pT$ here). As in Example 3, the inequalities here are meant componentwise.

The information transfers are schematized in Figure 4. The agents are denoted by Arabic numbers, and the Roman numbers describe the different stages of the spanning completion. Only the communications with guaranteed coefficient γ are represented. For simplicity, the self-loops are omitted.

Analyzing the system with the use of a unique spanning-tree (Theorem 10) yields $\tilde{\rho}_1 = \left(1 - \gamma^3\right)^{1/5}$.

To use Theorem 16 for analysis, one considers two spanning-trees and takes on any interval $[t_p, t_{p+1}]$ as follows:

$$\mathcal{N}^1_{p,0} = \mathcal{N}^2_{p,0} = \{1\}, \quad \mathcal{N}^1_{p,1} = \mathcal{N}^2_{p,1} = \{1, 2\},$$
$$\mathcal{N}^1_{p,2} = \{1, 2, 3\},\ \mathcal{N}^2_{p,2} = \{1, 2, 4\}, \quad \mathcal{N}^1_{p,3} = \{1, 2, 3, 5\},\ \mathcal{N}^2_{p,3} = \{1, 2, 4, 6\},$$
$$\mathcal{N}^1_{p,4} = \{1, 2, 3, 5, 6\},\ \mathcal{N}^2_{p,4} = \{1, 2, 4, 5, 6\}, \quad \mathcal{N}^1_{p,5} = \mathcal{N}^2_{p,5} = \mathcal{N}$$

and

$$A(0) = \cdots = A(4) = \begin{pmatrix} \gamma & 0 \\ 0 & \gamma \end{pmatrix}.$$

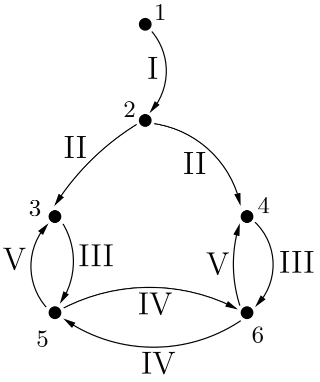


FIG. 4. *Schematized information transfer; see Example* 8.

The deduced estimate is $\tilde{\rho}_2 = \left(1-\gamma^5\right)^{1/5}$. The important point is that it is systematically *looser* than the previous one. Indeed, the previous formula could have been obtained by taking into account only one of the two spanning-trees, say, the "right branch," where the signal circulates in the order 1–2–4–6–5–3. In other words, there would be no difference in evaluating the graph similarly schematized, as shown in Figure 5.

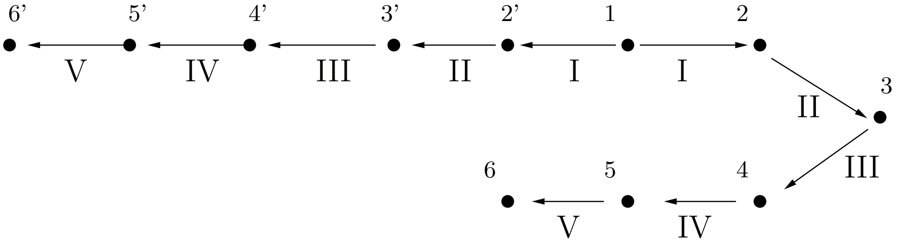


FIG. 5. *Schematized information transfer; see Example* 8.

How are we to take into account the crossing of the two spanning-trees, which is a case explicitly discarded in section 4.3? A general idea is to introduce new "populations." However, this is not so easy, as Lemma 18 is hardly adapted to this case (here it is useful to recall that the diameter of the union of two sets is at most equal to the sum of the diameters of these two sets *if their intersection is nonvoid*).

Along these lines, one may propose an idea for improvement of $\tilde{\rho}_2$. Let

$$\mathcal{N}^3_{p,t} \doteq \mathcal{N}^1_{p,t} \cup \mathcal{N}^2_{p,t} \ .$$

This is, in fact, just the population considered in the one-spanning-tree method leading to $\tilde{\rho}_1$. We are then allowed to take

$$A(0) = \cdots = A(2) = \begin{pmatrix} \gamma & 0 & 0 \\ 0 & \gamma & 0 \\ 0 & 0 & \gamma \end{pmatrix},\ A(3) = A(4) = \begin{pmatrix} \gamma & 0 & 1-\gamma \\ 0 & \gamma & 1-\gamma \\ 0 & 0 & \gamma \end{pmatrix} \ . \tag{18}$$

As

$$A(4)A(3)A(2)A(1)A(0) = \gamma^4 \begin{pmatrix} \gamma & 0 & 2(1-\gamma) \\ 0 & \gamma & 2(1-\gamma) \\ 0 & 0 & \gamma \end{pmatrix} ,$$

the estimate obtained via Theorem 16 is

$$\tilde{\rho}_3 \doteq \left(1 - \gamma^4(2-\gamma)\right)^{1/5} ,$$

which verifies $\tilde{\rho}_1 \leq \tilde{\rho}_3 \leq \tilde{\rho}_2$ for $\gamma \in [0,1]$; actually, $\tilde{\rho}_3$ does not overpass the precision of $\tilde{\rho}_1$. □

A careful examination of the previous example shows why no improvement could be obtained; i.e., the diameters of the three sets are equal up to the third stage, and the form of the difference inequalities involved forbid the two components fed by the third one to become larger than the latter.

However, notice that this paradoxical behavior is also a result of the value of the coefficients. The next example indicates that the method proposed in Example 8 can indeed provide better estimates.

Example 9. We consider a slight modification of Example 8. For fixed $\eta \in [0,1]$, we take Γ as previously, except

$$\Gamma(pT+1) \geq \begin{pmatrix} \gamma & 0 & 0 & 0 & 0 & 0 \\ 0 & \gamma & 0 & 0 & 0 & 0 \\ 0 & \eta\gamma & 0 & 0 & 0 & 0 \\ 0 & \gamma & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} , \quad \Gamma(pT+2) \geq \begin{pmatrix} \gamma & 0 & 0 & 0 & 0 & 0 \\ 0 & \gamma & 0 & 0 & 0 & 0 \\ 0 & 0 & \eta\gamma & 0 & 0 & 0 \\ 0 & 0 & 0 & \gamma & 0 & 0 \\ 0 & 0 & \gamma & 0 & 0 & 0 \\ 0 & 0 & 0 & \gamma & 0 & 0 \end{pmatrix} .$$

In other words, the transmission along the "left branch" in Figure 4 occurs with a smaller coefficient than along the right one. This modifies the evolution of the diameters of both $\mathcal{N}^1$ and $\mathcal{N}^3$, and one now has to modify the values of A by taking

$$A(1) = A(2) = \begin{pmatrix} \eta\gamma & 0 & 0 \\ 0 & \gamma & 0 \\ 0 & 0 & \eta\gamma \end{pmatrix} ,$$

instead of those given in (18). Using the notation of Example 8 yields the two contraction rate estimates

$$\tilde{\rho}_1' \doteq \left(1 - \eta^2\gamma^3\right)^{1/5} \quad \text{and} \quad \tilde{\rho}_3' \doteq \left(1 - \gamma^4(\gamma + 2\eta^2(1-\gamma))\right)^{1/5} .$$

In particular, when

$$\eta \leq \frac{\gamma}{(1+2\gamma(1-\gamma))^{1/2}}$$

(a quantity located in $[0,1/3]$ for $\gamma \in [0,1/2]$), the $\tilde{\rho}_3'$ is smaller than the estimate $\tilde{\rho}_1'$, obtained by considering a single spanning-tree. □

5. Conclusion. Several tools for estimating the convergence rate to consensus in multiagent systems were introduced and illustrated through simple examples. The criteria are based on topological as well as basic quantitative information. In accordance with previous results, consensus is reached provided that information can flow at least along some spanning-tree from one agent to all of the others. A key quantity, in this respect, appears to be a lower-bound on the total weight of the agents located upstream along the information flow for any chosen spanning-tree. More general criteria are also provided in which tighter estimates are allowed, provided that more spanning-trees are simultaneously taken into account.

These techniques are, in general, based on the idea of considering a decomposition of the overall population into subsets which influence each other in some quantifiable ways. Natural candidates for this partition appear to be the agents already attained by the information flows along the spanning-trees. There seem to be technical difficulties in trying to consider other kinds of partitions as, in general, neither the diameter of a union of sets nor the diameter of an intersection of sets is related to the diameters of the two sets. However, it may be possible to first consider the set of agents attained by one or more spanning-trees and then the set of agents attained in the reverse order. We leave this as an interesting open question for future research.

The method presented here provides results which are rather tight and inherently robust due to the qualitative nature of the assumptions involved. It is especially interesting to develop tools for quantitative estimates based on the consideration of simultaneous trees as arising from a single tree which gets repeated through time, as in Example 7. Again this will be a topic of future investigations.

Appendix A. Fundamental inequalities. We state in what follows a result on difference inequalities which is central to the techniques developed in the text. Consider the time-varying linear system (2). As before, the index set $\mathcal{N}$ is finite or countable, the x_k constitute a collection of scalar functions defined on $\mathbb{N}$, $\mathcal{I}$ is a finite or countable index set, and, for any $t \in \mathbb{N}$, a collection of subsets $\mathcal{M}_i(t)$ of $\mathcal{N}$, $i \in \mathcal{I}$, is given. Also, the state matrices $(\gamma_{k,l}(t))_{(k,l)\in\mathcal{N}\times\mathcal{N}}$ of the system are row-stochastic. Define the diameters

$$\Delta(t) \doteq \sup_{k\in\mathcal{N}} x_k(t) - \inf_{k\in\mathcal{N}} x_k(t), \qquad \Delta_i(t) \doteq \mathtt{diam}\, \mathcal{M}_i(t) = \sup_{k\in\mathcal{M}_i(t)} x_k(t) - \inf_{k\in\mathcal{M}_i(t)} x_k(t)$$

and the vector

$$\Delta_{\mathcal{M}}(t) \doteq (\mathtt{diam}\, \mathcal{M}_i(t))_{i\in\mathcal{I}} = (\Delta_i(t))_{i\in\mathcal{I}} \ .$$

The following result provides information on the evolution of the diameter vector.

Lemma 18. *Assume that for all $k \in \mathcal{N}$, for all $t \in \mathbb{N}$, some sets $\mathcal{M}_{k,j}(t)$, $j \in \mathcal{I}$, are given such that*

$$\mathcal{M}_{k,j}(t) \subseteq \mathcal{M}_j(t) \quad \text{and} \quad \mathcal{M}_{k,j}(t) \cap \mathcal{M}_{k,j'}(t) \neq \emptyset \Rightarrow j = j' \ .$$

Let maps $c_{i,j}(t)$, $i, j \in \mathcal{I}$, and $C(t)$ be such that

$$(19) \qquad c_{i,j}(t) \leq \inf_{k\in\mathcal{M}_i(t+1)} \sum_{l\in\mathcal{M}_{k,j}(t)} \gamma_{k,l}(t), \qquad C(t) \doteq (c_{i,j}(t))_{(i,j)\in\mathcal{I}\times\mathcal{I}} \ .$$

Then, for any $t, T \in \mathbb{N}$, for any $i \in \mathcal{I}$,

$$(20) \quad 0 \leq \Delta_i(t+T) \\ \leq \mathbf{1}_i^{\mathsf{T}} \Big(C(t+T-1)\dots C(t)\Delta_{\mathcal{M}}(t) + (\mathbf{1} - C(t+T-1)\dots C(t)\mathbf{1})\, \Delta(t) \Big) \ .$$

By convention, we put $\sum_{i\in\emptyset} c_i = 0$ and $\inf_{i\in\emptyset} c_i = +\infty$. Recall that the vector $\mathbf{1}$ in the statement is made up of a column of 1 and that the vector $\mathbf{1}_i$ has null components, except 1 in the ith position (in finite dimension, it is the ith vector of the canonic basis). In particular, $\mathbf{1}_i^\intercal \Delta_{\mathcal{M}}(t) = \Delta_i(t)$, $\mathbf{1} = \sum_{i\in\mathcal{I}} \mathbf{1}_i$.

Remark 5. Notice that formula (20) may involve infinite summations in the products of infinite-dimensional matrices. As the coefficients of the matrices $C(t)$ are nonnegative and bounded by 1, uniform convergence of the series of terms indeed occurs on any bounded time interval, and therefore the notation has a univocal meaning.

Proof. Define

$$M(t) \doteq \sup_{k\in\mathcal{N}} x_k(t),\ M_i(t) \doteq \sup_{k\in\mathcal{M}_i(t)} x_k(t),\ m(t) \doteq \inf_{k\in\mathcal{N}} x_k(t),\ m_i(t) \doteq \inf_{k\in\mathcal{M}_i(t)} x_k(t)$$

in such a way that the quantities previously defined in the statement verify

$$\Delta \equiv M - m,\ \Delta_i \equiv M_i - m_i\ .$$

First of all, notice that, due to the nonnegativity of the coefficients $\gamma_{k,l}(t)$, identity (2) implies, for any $t\in\mathbb{N}$ and for any $k\in\mathcal{N}$,

$$x_k(t+1) \leq \sum_{l\in\mathcal{N}} \gamma_{k,l}(t) M(t) = M(t)\ .$$

Taking the supremum and arguing similarly for the lower-bounds, we obtain

$$M(t+1) \leq M(t),\ m(t+1) \geq m(t)\ .$$

In particular,

$$\Delta(t+1) \leq \Delta(t)\ . \tag{21}$$

Also, due to the fact that $\mathcal{M}_i(t) \subseteq \mathcal{N}$, it holds that

$$M_i(t) = \sup_{k\in\mathcal{M}_i(t)} x_k(t) \leq \sup_{k\in\mathcal{N}} x_k(t) = M(t), \quad m_i(t) \geq m(t)\ , \tag{22}$$

and

$$\Delta_i(t) \leq \Delta(t)\ .$$

Applying a tighter estimate, one obtains from (2) that, for any $k\in\mathcal{N}$,

$$\begin{aligned} x_k(t+1) &= \sum_{j\in\mathcal{I}} \sum_{l\in\mathcal{M}_{k,j}(t)} \gamma_{k,l}(t) x_l(t) + \sum_{l\in\mathcal{N}\setminus\bigcup_{j\in\mathcal{I}}\mathcal{M}_{k,j}(t)} \gamma_{k,l}(t) x_l(t) \\ &\leq \sum_{j\in\mathcal{I}} \left(\sum_{l\in\mathcal{M}_{k,j}(t)} \gamma_{k,l}(t) \right) M_j(t) + \sum_{l\in\mathcal{N}\setminus\bigcup_{j\in\mathcal{I}}\mathcal{M}_{k,j}(t)} \gamma_{k,l}(t) M(t) \\ &= \sum_{j\in\mathcal{I}} \sum_{l\in\mathcal{M}_{k,j}(t)} \gamma_{k,l}(t) M_j(t) + \left(1 - \sum_{j\in\mathcal{I}} \sum_{l\in\mathcal{M}_{k,j}(t)} \gamma_{k,l}(t) \right) M(t)\ , \end{aligned}$$

due to (19) and the nonnegativity of the coefficients $\gamma_{k,l}(t)$. If now $k \in \mathcal{M}_i(t+1)$ for some $i \in \mathcal{I}$, one obtains

$$\begin{aligned} x_k(t+1) &\leq \sum_{j\in\mathcal{I}} c_{i,j}(t)M_j(t) + \left(1 - \sum_{j\in\mathcal{I}} c_{i,j}(t)\right) M(t) \\ &\quad + \sum_{j\in\mathcal{I}} \left(c_{i,j}(t) - \sum_{l\in\mathcal{M}_{k,j}(t)} \gamma_{k,l}(t)\right)(M(t) - M_j(t)) \\ &\leq \sum_{j\in\mathcal{I}} c_{i,j}(t)M_j(t) + \left(1 - \sum_{j\in\mathcal{I}} c_{i,j}(t)\right) M(t)\ , \end{aligned}$$

due to the fact that $M(t) \geq M_j(t)$ for any $t \in \mathbb{N}$ and any $j \in \mathcal{I}$ (see (22)). Consequently, for any $i \in \mathcal{I}$,

$$M_i(t+1) \leq \sum_{j\in\mathcal{I}} c_{i,j}(t)M_j(t) + \left(1 - \sum_{j\in\mathcal{I}} c_{i,j}(t)\right) M(t)\ .$$

One establishes similarly that

$$x_k(t+1) \geq \sum_{j\in\mathcal{I}} c_{i,j}(t)m_j(t) + \left(1 - \sum_{j\in\mathcal{I}} c_{i,j}(t)\right) m(t)\ ,$$

with the *same* coefficients, so, for any $i \in \mathcal{I}$,

$$m_i(t+1) \geq \sum_{j\in\mathcal{I}} c_{i,j}(t)m_j(t) + \left(1 - \sum_{j\in\mathcal{I}} c_{i,j}(t)\right) m(t)\ .$$

Subtracting the previous inequalities, one may thus deduce that, for any $i \in \mathcal{I}$,

$$\Delta_i(t+1) \leq \sum_{j\in\mathcal{I}} c_{i,j}(t)\Delta_j(t) + \left(1 - \sum_{j\in\mathcal{I}} c_{i,j}(t)\right) \Delta(t)\ .$$

The collection of these inequalities, together with (21), may be written under the matrix (possibly infinite) form

$$\begin{pmatrix} \Delta_{\mathcal{M}}(t+1) \\ \Delta(t+1) \end{pmatrix} \leq \begin{pmatrix} C(t) & \mathbf{1} - C(t)\mathbf{1} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \Delta_{\mathcal{M}}(t) \\ \Delta(t) \end{pmatrix}\ .$$

The previous inequality has to be understood componentwise.

Now, one shows easily that

$$\begin{aligned} &\begin{pmatrix} C(t+1) & \mathbf{1} - C(t+1)\mathbf{1} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} C(t) & \mathbf{1} - C(t)\mathbf{1} \\ 0 & 1 \end{pmatrix} \\ &\qquad\qquad = \begin{pmatrix} C(t+1)C(t) & \mathbf{1} - C(t+1)C(t)\mathbf{1} \\ 0 & 1 \end{pmatrix} \end{aligned}$$

in such a way that, for nonnegative T,

$$\begin{pmatrix}\Delta_{\mathcal{M}}(t+T)\\ \Delta(t+T)\end{pmatrix} \leq \begin{pmatrix}C(t+T-1)\dots C(t) & \mathbf{1}-C(t+T-1)\dots C(t)\mathbf{1}\\ 0 & 1\end{pmatrix}\begin{pmatrix}\Delta_{\mathcal{M}}(t)\\ \Delta(t)\end{pmatrix}\ .$$

This formula permits us to complete the proof of Lemma 18. □

Appendix B. Proof of Lemma 17.

1. One verifies directly that, for any $t = 0, \dots, n+q-3$, $A(t)$ is equal to

$$I_q - (1-\gamma)\Big(\delta_{0\leq t\leq n-1}\ e_1e_1^{\mathsf{T}} + \delta_{1\leq t\leq n}\ e_2(e_2-e_1)^{\mathsf{T}} + \cdots \\ + \delta_{q-1\leq t\leq n+q-2}\ e_q(e_q-e_{q-1})^{\mathsf{T}}\Big)\ ,$$

where e_i is the ith vector of the canonical basis in $\mathbb{R}^q$ and $\delta_{i\leq t\leq i+n-1}$ is 1 (resp., 0) if the condition written in the index is fulfilled (resp., violated).

Let us first establish the following factorization formula:

$$(23)\quad A(n+q-3)\dots A(0) \\ = \gamma^{n-q}\ \mathtt{diag}\{1;\gamma;\dots;\gamma^{q-1}\}\ B(n+q-3)\dots B(0)\ \mathtt{diag}\{\gamma^{q-1};\dots;\gamma;1\}\ ,$$

where the matrix $B(t)$ is obtained from $A(t)$ by replacing γ on the diagonal by 1, and $1-\gamma$ by

$$\xi \doteq \frac{1-\gamma}{\gamma}\ ,$$

that is, simply,

$$B(t) \doteq I_q + \xi\sum_{i=1}^{q-1}\delta_{i\leq t\leq i+n-2}\ e_{i+1}e_i^{\mathsf{T}}\ .$$

In formula (23) and below, `diag` is used to define diagonal matrices.

Formula (23) will be proved by induction on the positive integer q. Notice that, strictly speaking, the matrices $A, B \in \mathbb{R}^{q\times q}$ depend upon q (and n), but for simplicity we omit here any explicit indication of this dependence. Indeed, for $q = 1$, $A(t) = \gamma$ for $0 \leq t \leq n+q-3 = n-2$, and $A(n-2)\dots A(0) = \gamma^{n-1}$, while, for $q = 2$, $n+q-3 = n-1$ and

$$A(0) = \begin{pmatrix}\gamma & 0\\ 0 & 1\end{pmatrix},\quad A(t) = \begin{pmatrix}\gamma & 0\\ 1-\gamma & \gamma\end{pmatrix},\ 1\leq t\leq n-2,\quad A(n-1) = \begin{pmatrix}1 & 0\\ 1-\gamma & \gamma\end{pmatrix}\ ,$$

so that

$$A(n-1)\dots A(0) = \gamma^{n-2}\begin{pmatrix}1 & 0\\ 0 & \gamma\end{pmatrix}\begin{pmatrix}1 & 0\\ \xi & 1\end{pmatrix}^{n-1}\begin{pmatrix}\gamma & 0\\ 0 & 1\end{pmatrix} \\ = \gamma^{n-2}\begin{pmatrix}1 & 0\\ 0 & \gamma\end{pmatrix}B(n-1)\dots B(0)\begin{pmatrix}\gamma & 0\\ 0 & 1\end{pmatrix}\ .$$

(Notice that $B(0) = I_2$ and $B(t) = \left(\begin{smallmatrix}1 & 0\\ \xi & 1\end{smallmatrix}\right)$ for $t = 1, \dots, n-1$.)

Assume now that (23) is true at order $q-1$ and consider order q. Due to the particular structure of the matrices A and B, which are null except for the terms on the diagonal and the subdiagonal, one has

(24a)

$$\mathtt{diag}\{I_{q-1};0\}\left(\prod_{t=0}^{n+q-3} A(t)\right)\mathtt{diag}\{I_{q-1};0\} = \prod_{t=0}^{n+q-3} \mathtt{diag}\{I_{q-1};0\}A(t)\ \mathtt{diag}\{I_q;0\},$$

(24b)

$$\mathtt{diag}\{0;I_{q-1}\}\left(\prod_{t=0}^{n+q-3} A(t)\right)\mathtt{diag}\{0;I_{q-1}\} = \prod_{t=0}^{n+q-3} \mathtt{diag}\{0;I_{q-1}\}A(t)\ \mathtt{diag}\{0;I_{q-1}\},$$

and similarly for $B(t)$. In the previous identities and in subsequent formulas, the products are noncommutative: the convention is that t is decreasing from the left factor to the right one.

Now, it is easy to identify the right-hand sides of the two identities (24) with a product of $n+(q-1)-3 = n+q-4$ matrices A (resp., B) corresponding to the index $q-1$ (the last term in the right-hand product in (24a), resp., the first term in the right-hand product in (24b), is equal to $\mathtt{diag}\{I_{q-1};0\}$, resp., $\mathtt{diag}\{0;I_{q-1}\}$, and can be suppressed). Using the induction hypothesis at order $q-1$, one shows that

$$\begin{aligned}
&\mathtt{diag}\{I_{q-1};0\}\left(\prod_{t=0}^{n+q-3} A(t)\right)\mathtt{diag}\{I_{q-1};0\}\\
&= \gamma^{n-(q-1)}\ \mathtt{diag}\{1;\dots;\gamma^{q-2};0\}\ \mathtt{diag}\{I_{q-1};0\}\ \left(\prod_{t=0}^{n+q-3} B(t)\right)\\
&\qquad \cdot \mathtt{diag}\{I_{q-1};0\}\ \mathtt{diag}\{\gamma^{q-2};\dots;1;0\}\\
&= \gamma^{n-(q-1)}\gamma^{-1}\ \mathtt{diag}\{I_{q-1};0\}\ \mathtt{diag}\{1;\dots;\gamma^{q-1}\}\ \left(\prod_{t=0}^{n+q-3} B(t)\right)\\
&\qquad \cdot \mathtt{diag}\{\gamma^{q-1};\dots;1\}\ \mathtt{diag}\{I_{q-1};0\}\\
&= \gamma^{n-q}\ \mathtt{diag}\{I_{q-1};0\}\ \mathtt{diag}\{1;\dots;\gamma^{q-1}\}\ \left(\prod_{t=0}^{n+q-3} B(t)\right)\\
&\qquad \cdot \mathtt{diag}\{\gamma^{q-1};\dots;1\}\ \mathtt{diag}\{I_{q-1};0\}.
\end{aligned}$$

One establishes similarly that

$$\begin{aligned}
&\mathtt{diag}\{0;I_{q-1}\}\left(\prod_{t=0}^{n+q-3} A(t)\right)\mathtt{diag}\{0;I_{q-1}\}\\
&\quad = \gamma^{n-q}\ \mathtt{diag}\{0;I_{q-1}\}\ \mathtt{diag}\{1;\dots;\gamma^{q-1}\}\ \left(\prod_{t=0}^{n+q-3} B(t)\right)\\
&\qquad\qquad \cdot\ \mathtt{diag}\{\gamma^{q-1};\dots;1\}\ \mathtt{diag}\{0;I_{q-1}\}\ ,
\end{aligned}$$

and this is indeed sufficient, due to the structure of the matrices A and B mentioned earlier, to prove that (23) is true at order q. This achieves the proof of (23) by induction.

2. One now estimates the matrix-product

$$\Pi = (\Pi_{i,j})_{(i,j)\in\{1,\dots,q\}^2} \doteq \prod_{t=0}^{n+q-3} B(t) = \prod_{t=1}^{n+q-3} \left(I_q + \xi \sum_{i=1}^{q-1} \delta_{i\leq t\leq i+n-2}\ e_{i+1}e_i^\intercal \right) .$$

Each term of this product is a lower-triangular matrix, so Π shares the same property.

The fact that the canonical basis is orthonormal implies that, for any $i > j$, $i,j \in \{1,\dots,q\}$, it holds that

$$\Pi_{i,j} \doteq \xi^{i-j} \text{ card} \Big\{ (t_{j+1},\dots,t_i) \in [j, j+n-2] \times \dots \times [i-1, i+n-3] \cap \mathbb{N}^{i-j} \ : \ t_{j+1} < \dots < t_i \Big\} ,$$

and also that the diagonal terms are equal to 1. The previous formula just means that, for a term in $e_ie_j^\intercal$ to emerge from the product, it should be the result of the product

$$(e_ie_{i-1}^\intercal)\cdot(e_{i-1}e_{i-2}^\intercal)\dots(e_{j+1}e_j^\intercal) ,$$

where each of the terms in parentheses comes from a certain matrix $A(t)$—the rest of the factors come from identity matrices. Conversely, all products of different type vanish.

In order to evaluate the quantities $\Pi_{i,j}$ previously defined, notice that the change of variables

$$t'_{j+1} = t_{j+1},\ t'_{j+2} = t_{j+2} - 1, \dots, t'_i = t_i - (i-j-1)$$

yields

$$\text{card} \Big\{ (t_{j+1},\dots,t_i) \in [j, j+n-2] \times \dots \times [i-1, i+n-3] \cap \mathbb{N}^{i-j} \ : \ t_{j+1} < \dots < t_i \Big\} = \text{card} \Big\{ (t'_{j+1},\dots,t'_i) \in ([j, j+n-2] \cap \mathbb{N})^{i-j} \ : \ t'_{j+1} \leq \dots \leq t'_i \Big\} .$$

3. We now compute explicitly the value of the function $F(m,n)$ defined on $\mathbb{N}\times\mathbb{N}$ as

$$F(m,n) \doteq \text{card} \Big\{ (t_1,\dots,t_m) \in ([1,n] \cap \mathbb{N})^m \ : \ t_1 \leq \dots \leq t_m \Big\} .$$

Clearly,

$$F(1,n) = n, \quad F(2,n) = \frac{n(n+1)}{2} .$$

Considering separately the cases where $t_1 = 1$, $t_1 = 2, \dots, t_1 = n$, one finds the following induction relation:

$$F(m,n) = \sum_{i=1}^{n} F(m-1,i) .$$

On the other hand, let

$$G(m,n) \doteq \binom{m+n-1}{m} = \frac{(m+n-1)!}{m!(n-1)!} \ ;$$

one has

$$G(1,n) = \binom{n}{1} = n, \quad G(2,n) = \binom{n+1}{2} = \frac{n(n+1)}{2} \ .$$

Independently, it is known that

$$\binom{n}{m} = \binom{n-1}{m-1} + \binom{n-1}{m}$$

in such a way that

$$G(m,n) = \binom{m+n-1}{m} = \binom{m+n-2}{m-1} + \binom{m+n-2}{m} = G(m-1,n)+G(m,n-1) \, .$$

It ensues, from repeated use of this formula, that

$$\begin{aligned} G(m,n) = G(m-1,n)+G(m,n-1) &= G(m-1,n)+G(m-1,n-2)+G(m,n-3) \\ = \dots = \sum_{i=2}^{n} G(m-1,i) + G(m,1) &= \sum_{i=1}^{n} G(m-1,i) \ , \end{aligned}$$

because $G(m,1) = G(m-1,1) = 1$. Having the same initial condition and sharing the same induction relation, F and G are thus equal, and $F(m,n) = \binom{m+n-1}{m}$.

4. The value of F found before is now used to estimate first Π and then $\tilde{\rho}_q$. We deduce from what precedes that, for $i > j$,

$$\Pi_{i,j} = \xi^{i-j} F(i-j,n-1) = \xi^{i-j} \binom{i-j+n-2}{i-j} = \xi^{i-j} \binom{i-j+n-2}{n-2} \ .$$

Recall that $\Pi_{i,i} = 1$ and $\Pi_{i,j} = 0$ for $i < j$.

From the fact that the matrix Π above is lower-triangular, one finds by application of Theorem 16 that

$$1 - \tilde{\rho}_q^{n+q-2} = \max_{i=1,\dots,q} \mathbf{1}_i A(n+q-3)\dots A(0)\mathbf{1} = \mathbf{1}_q A(n+q-3)\dots A(0)\mathbf{1} \ .$$

From (23) and the previous computations, one thus deduces

$$\begin{aligned} 1 - \tilde{\rho}_q^{n+q-2} = \gamma^{n-1} \sum_{i=1}^{q} \gamma^{q-i} \Pi_{q,i} &= \gamma^{n-1} \sum_{i=1}^{q} \gamma^{q-i} \xi^{q-i} \binom{q-i+n-2}{n-2} \\ &= \gamma^{n-1} \sum_{i=1}^{q} \gamma^{q-i} \left(\frac{1-\gamma}{\gamma}\right)^{q-i} \binom{q-i+n-2}{n-2} \ . \end{aligned}$$

Thus,

$$\tilde{\rho}_q^{n+q-2} = 1 - \gamma^{n-1} \sum_{i=1}^{q} (1-\gamma)^{q-i} \binom{q-i+n-2}{n-2} \ ,$$

or again,

$$\tilde{\rho}_q = \left(1 - \gamma^{n-1} \sum_{i=0}^{q-1} (1-\gamma)^i \begin{pmatrix} n+i-2 \\ n-2 \end{pmatrix} \right)^{1/(n+q-2)} .$$

This achieves the proof of the first equality in the statement of Lemma 17.

5. To show the identity of the two expressions in Lemma 17, notice that

$$\sum_{i=0}^{q-1} \begin{pmatrix} n+i-2 \\ n-2 \end{pmatrix} \delta^i = \frac{1}{(n-2)!} \sum_{i=0}^{q-1} \frac{d^{n-2}}{d\delta^{n-2}} \left[\delta^{n+i-2} \right] = \frac{1}{(n-2)!} \frac{d^{n-2}}{d\delta^{n-2}} \left[\sum_{i=0}^{q-1} \delta^{n+i-2} \right]$$
$$= \frac{1}{(n-2)!} \frac{d^{n-2}}{d\delta^{n-2}} \left[\frac{1-\delta^{n+q-2}}{1-\delta} \right] .$$

On the other hand, one shows easily that

$$\frac{1}{(n-2)!} \frac{d^{n-2}}{d\delta^{n-2}} \left[\frac{1}{1-\delta} \right] = \frac{1}{(1-\delta)^{n-1}} .$$

This permits us to deduce the identity of the two expressions in the statement.

6. Lastly, we show the limiting property expressed in Lemma 17. From the last formula, one may see that, for every $n \geq 2$,

$$\tilde{\rho}_q = \sqrt[n+q-2]{P_n(q,\delta)\delta^q} ,$$

where P_n is a polynomial in q and $\delta = 1-\gamma$ of degree $n-2$ with respect to both variables. Henceforth, taking the limit for $q \to +\infty$ yields the estimate

$$\lim_{q \to +\infty} \tilde{\rho}_q = \lim_{q \to +\infty} \sqrt[n+q-2]{P_n(q,\delta)\delta^q} = \delta = 1-\gamma ,$$

which corresponds to the true value of the converging rate. Indeed,

$$\sqrt[n+q-2]{P_n(q,\delta)} = e^{\ln P_n(q,\delta)/(n+q-2)} = e^{[(n-2)\ln q + \ln(1+O(1/q))]/(n+q-2)} ,$$

as P_n is of degree $n-2$ in q. The asymptotic expansion announced in the statement is thus proved, and this achieves the proof of Lemma 17.

Acknowledgments. The authors wish to thank the reviewers and the associate editor for their penetrating comments, including the detection of a mistake in the previous version.

REFERENCES

[1] M. Akar and R. Shorten, *Time synchronization for wireless sensors networks*, in Proceedings of the 17th Symposium on Mathematical Theory of Networks and Systems, Kyoto, Japan, 2006, pp. 221–225.

[2] H. Ando, Y. Oasa, I. Suzuki, and M. Yamashita, *Distributed memoryless point convergence algorithm for mobile robots with limited visibility*, IEEE Trans. Robotics and Automation, 15 (1999), pp. 818–828.

[3] D. Angeli and P.-A. Bliman, *Stability of leaderless discrete-time multi-agent systems*, Math. Control Signals Systems, 18 (2006), pp. 293–322.

[4] D. Angeli and P.-A. Bliman, *Convergence Speed of Unsteady Distributed Consensus: Decay Estimate Along the Settling Spanning-Trees*, http://arxiv.org/abs/math.OC/0610854 (2006).

[5] A. Berman, T. Laffey, T. Leizarowitz, and R. Shorten, *On the second eigenvalues of matrices associated with TCP*, Linear Algebra Appl., 416 (2006), pp. 175–183.

[6] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation*, Prentice–Hall International, London, 1989; also available online at https://dspace.mit.edu/handle/1721.1/3719.

[7] V. D. Blondel, J. M. Hendrickx, A. Olshevsky, and J. N. Tsitsiklis, *Convergence in multiagent coordination, consensus, and flocking*, in Proceedings of the Joint European Control Conference/44th Annual IEEE Conference on Decision and Control, Sevilla, Spain, 2005, pp. 2996–3000.

[8] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, *Randomized gossip algorithms*, IEEE Trans. Inform. Theory (special issue of IEEE Trans. Inform. Theory and IEEE/ACM Trans. Networking), 52 (2006), pp. 2508–2530.

[9] M. Cao, A. S. Morse, and B. D. O. Anderson, *Reaching a consensus in a dynamically changing environment: Convergence rates, measurement delays, and asynchronous events*, SIAM J. Control Optim., 47 (2008), pp. 601–623.

[10] M. Cao, D. A. Spielman, and A. S. Morse, *A lower bound on convergence of a distributed network consensus algorithm*, in Proceedings of the Joint European Control Conference/44th Annual IEEE Conference on Decision and Control, Sevilla, Spain, 2005, pp. 2356–2361.

[11] J. Cheeger, *A lower bound for the smallest eigenvalue of the Laplacian*, in Problems in Analysis. Papers Dedicated to Salomon Bochner, Princeton University Press, Princeton, NJ, 1969, pp. 195–199.

[12] H. Cohn, *Products of stochastic matrices and applications*, Int. J. Math. Math. Sci., 12 (1989), pp. 209–233.

[13] A. Czirók and T. Vicsek, *Collective motion*, in Statistical Mechanics of Biocomplexity, D. Reguera, M. Rubi, and J. Vilar, eds., Lecture Notes in Phys. 527, Springer-Verlag, Berlin, 1999, pp. 152–164.

[14] P. Diaconis and D. Stroock, *Geometric bounds for eigenvalues of Markov chains*, Ann. Appl. Probab., 1 (1991), pp. 36–61.

[15] S. Friedland and R. Nabben, *On the second real eigenvalue of nonnegative and Z-matrices*, Linear Algebra Appl., 255 (1997), pp. 303–313.

[16] P. M. Gade, *Feedback control in coupled map lattices*, Phys. Rev. E, 57 (1998), pp. 7309–7312.

[17] S. T. Garren and R. L. Smith, *Estimating the second largest eigenvalue of a Markov transition matrix*, Bernoulli, 6 (2000), pp. 215–242.

[18] O. Gross and U. G. Rothblum, *Approximations of the spectral radius, corresponding eigenvector, and second largest modulus of an eigenvalue for square, nonnegative, irreducible matrices*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 15–32.

[19] Y. Hatano and M. Mesbahi, *Agreement over random networks*, IEEE Trans. Automat. Control, 50 (2005), pp. 1867–1872.

[20] M. W. Hirsch, *Convergent activation dynamics in continuous time networks*, Neural Networks, 2 (1989), pp. 331–349.

[21] A. Jadbabaie, N. Motee, and M. Barahona, *On the stability of the Kuramoto model of coupled nonlinear oscillators*, in Proceedings of the IEEE American Control Conference, Boston, MA, 2004, pp. 4296–4301.

[22] H. J. Landau and A. M. Odlyzko, *Bounds for eigenvalues of certain stochastic matrices*, Linear Algebra Appl., 38 (1981), pp. 5–15.

[23] N. E. Leonard and E. Fiorelli, *Virtual leaders, artificial potentials, and coordinated control of groups*, in Proceedings of the 40th Annual IEEE Conference on Decision and Control, Orlando, FL, 2001, pp. 2968–2973.

[24] J. Lorenz, *Consensus strikes back in the Hegselmann–Krause model of continuous opinion dynamics under bounded confidence*, J. Artificial Societies and Social Simulation, 9 (2006); available online at http://jasss.soc.surrey.ac.uk/9/1/8.html

[25] S. C. Manrubia and A. S. Mikhailov, *Mutual synchronization and clustering in randomly coupled chaotic dynamical networks*, Phys. Rev. E, 60 (1999), pp. 1579–1589; also available online at http://arxiv.org/abs/cond-mat/9905083.

[26] L. Moreau, *Time-Dependent Unidirectional Communication in Multi-Agent Systems*, http://arxiv.org/abs/math/0306426 (2003).

[27] L. Moreau, *A note on leaderless communication via bidirectional and unidirectional time-dependent communication*, in Proceedings of the MTNS 2004, Leuven, Belgium, 2004.

[28] L. Moreau, *Stability of multi-agent systems with time-dependent communication links*, IEEE Trans. Automat. Control, 50 (2005), pp. 169–182.

[29] A. Nedich and A. Ozdaglar, *Convergence Rate for Consensus with Delays*, LIDS Technical Report 2774, Laboratory for Information and Decision Systems, MIT, Boston, MA, 2007 (submitted for publication).

[30] A. Olshevsky and J. N. Tsitsiklis, *Convergence speed in distributed consensus and averaging*, in Proceedings of the 45th IEEE Conference on Decision and Control, San Diego, CA, 2006, pp. 3387–3392.

[31] R. Sepulchre, D. Paley, and N. Leonard, *Collective motion and oscillator synchronization*, in Proceedings of the 2003 Block Island Workshop on Cooperative Control, Springer-Verlag, New York, 2005, pp. 189–228.

[32] J. N. Tsitsiklis, *Problems in Decentralized Decision Making and Computation*, Ph.D. thesis, Department of Electrical Engineering and Computer Science, MIT, Boston, MA, also available online at http://hdl.handle.net/1721.1/15254, 1984.

[33] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans, *Distributed asynchronous deterministic and stochastic gradient optimization algorithms*, IEEE Trans. Automat. Control, 31 (1986), pp. 803–812.

[34] L. Yang, K. M. Passino, and M. M. Polycarpou, *Stability analysis of m-dimensional asynchronous swarms with a fixed communication topology*, IEEE Trans. Automat. Control, 48 (2003), pp. 76–95.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 1, pp. 33–55

© 2009 Society for Industrial and Applied Mathematics

CONVERGENCE SPEED IN DISTRIBUTED CONSENSUS AND AVERAGING[*]

ALEX OLSHEVSKY[†] AND JOHN N. TSITSIKLIS[†]

Abstract. We study the convergence speed of distributed iterative algorithms for the consensus and averaging problems, with emphasis on the latter. We first consider the case of a fixed communication topology. We show that a simple adaptation of a consensus algorithm leads to an averaging algorithm. We prove lower bounds on the worst-case convergence time for various classes of linear, time-invariant, distributed consensus methods, and provide an algorithm that essentially matches those lower bounds. We then consider the case of a time-varying topology, and provide a polynomial-time averaging algorithm.

Key words. consensus algorithms, distributed averaging, cooperative control

AMS subject classification. 93A14

DOI. 10.1137/060678324

1. Introduction.

Given a set of autonomous agents—which may be sensors, nodes of a communication network, cars, or unmanned aerial vehicles—the distributed *consensus* problem asks for a distributed algorithm that the agents can use to agree on an opinion (represented by a scalar or a vector), starting from different initial opinions among the agents, and in the presence of possibly severely restricted communications.

Algorithms that solve the distributed consensus problem provide the means by which networks of agents can be coordinated. Although each agent may have access to different local information, the agents can agree on a decision (e.g., on a common direction of motion, on the time to execute a move, etc.). Such synchronized behavior has often been observed in biological systems [15].

The distributed consensus problem has historically appeared in many diverse areas, such as parallel computation [30, 31, 3], control theory [18, 28], and communication networks [24, 22]. Recently, the problem has attracted significant attention [18, 22, 2, 11, 24, 7, 14, 25, 26, 13, 8, 5, 1] motivated by new contexts and open problems in communications, sensor networks, and networked control theory. We briefly describe some of the more recent applications.

Reputation management in ad hoc networks. It is often the case that the nodes of a wireless multihop network are not controlled by a single authority or do not have a common objective. Selfish behavior among nodes (e.g., refusing to forward traffic meant for others) is possible, and some mechanism is needed to enforce cooperation. One way to detect selfish behavior is reputation management; i.e., each node forms an opinion by observing the behavior of its neighbors. One is then faced with the problem of combining these different opinions into a single globally available reputation measure for each node. The use of distributed consensus algorithms for

[*] Received by the editors December 21, 2006; accepted for publication (in revised form) May 22, 2008; published electronically February 11, 2009. This research was supported by the National Science Foundation under a Graduate Research Fellowship and grants ECS-0312921 and ECS-0426453. A preliminary version of this paper was presented at the 45th IEEE Conference on Decision and Control, San Diego, CA, 2006.

http://www.siam.org/journals/sicon/48-1/67832.html

[†] Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139 (alex_o@mit.edu, jnt@mit.edu).

doing this was explored in [22], where a variation of one of the methods we examine here—the "agreement algorithm"—was used as a basis for an empirical investigation.

Sensor networks. A sensor network designed for detection or estimation needs to combine various measurements into a decision or into a single estimate. Distributed computation of this decision/estimate has the advantage of being fault-tolerant (network operation is not dependent on a small set of nodes) and self-organizing (network functionality does not require constant supervision) [31, 2, 11].

Control of autonomous agents. It is often necessary to coordinate collections of autonomous agents (e.g., cars or unmanned aerial vehicles). For example, one may wish for the agents to agree on a direction or speed. Even though the data related to the decision may be distributed through the network, it is usually desirable that the final decision depend on all the known data, even though most of the data are unavailable at each node. A model motivated by such a context was empirically investigated in [32].

In this paper, we focus on a special case of the distributed consensus problem, the distributed *averaging* problem. Averaging algorithms guarantee that the final global value will be the exact average of the initial individual values. Our general objective is to characterize the worst-case convergence time of various averaging algorithms, as a function of the number n of agents, and to understand their fundamental limitations by providing lower bounds on the convergence time.

We now outline the remainder of this paper and preview the main contributions. In section 2, we provide some background material by reviewing the agreement algorithm of [30, 31] for the distributed consensus problem. In sections 3–8, we consider the case of fixed graphs. In section 3, we discuss three different ways that the agreement algorithm can provide a solution to the averaging problem. In particular, we show how an averaging algorithm can be constructed based on two parallel executions of the agreement algorithm. In section 4, we define the notions of convergence rate and convergence time, and we provide a variational characterization of the convergence rate.

In section 5, we use results from [23] to show that the worst-case convergence time of an averaging algorithm introduced in section 3 is essentially $\Theta(n^3)$.[1] In section 6, we show that for one of our methods, the convergence rate can be made arbitrarily fast. On the other hand, under an additional restriction that reflects numerical stability considerations, we show that the convergence time of a certain class of algorithms (and by extension of a certain class of averaging algorithms) is $\Omega(n^2)$, in the worst case. We also provide a simple method (based on executing the agreement algorithm on a spanning tree) whose convergence time essentially matches the $\Omega(n^2)$ lower bound. In section 7, we discuss briefly particular methods that employ doubly stochastic matrices and their potential drawbacks.

Then, in section 8, we turn our attention to the case of dynamic topologies. For the agreement algorithm, we show that its convergence time for the case of nonsymmetric topologies can be exponentially large in the worst case. On the other hand, for the case of symmetric topologies, we provide a new averaging algorithm (and therefore, an agreement algorithm as well), whose convergence time is $O(n^3)$. To the best of our knowledge, none of the existing consensus or averaging algorithms

[1] Let f and g be two positive functions on the positive integers. We write $f(n) = O(g(n))$ (respectively, $f(n) = \Omega(g(n))$) if there exists a positive constant c and some n_0 such that $f(n) \leq cg(n)$ (respectively, $f(n) \geq cg(n)$) for all $n \geq n_0$. If $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$ both hold, we write $f(n) = \Theta(g(n))$.

in the literature has a similar guarantee of polynomial time convergence in the presence of dynamically changing topologies. In section 9, we report on some numerical experiments illustrating the advantages of two of our algorithms. Section 10 contains some brief concluding remarks.

2. The agreement algorithm. The "agreement algorithm" is an iterative procedure for the solution of the distributed consensus problem. It was introduced in [10] for the time-invariant case, and in [30, 31] for the case of "asynchronous" and time-varying environments. We briefly review this algorithm and summarize the available convergence results.

Consider a set $\mathcal{N} = \{1, 2, \ldots, n\}$ of nodes. Each node i starts with a scalar value $x_i(0)$; the vector with the values of all nodes at time t is denoted by $x(t) = (x_1(t), \ldots, x_n(t))$. The agreement algorithm updates $x(t)$ according to the equation $x(t+1) = A(t)x(t)$, or

$$x_i(t+1) = \sum_{j=1}^{n} a_{ij}(t)x_j(t),$$

where $A(t)$ is a nonnegative matrix with entries $a_{ij}(t)$. The row-sums of $A(t)$ are equal to 1, so that $A(t)$ is a stochastic matrix. In particular, $x_i(t+1)$ is a weighted average of the values $x_j(t)$ held by the nodes at time t.

We next state some conditions under which the agreement algorithm is guaranteed to converge.

ASSUMPTION 2.1. *There exists a positive constant α such that*

(a) $a_{ii}(t) \geq \alpha$ *for all i, t.*

(b) $a_{ij}(t) \in \{0\} \cup [\alpha, 1]$ *for all i, j, t.*

(c) $\sum_{j=1}^{n} a_{ij}(t) = 1$ *for all i, t.*

Intuitively, whenever $a_{ij}(t) > 0$, node j communicates its current value $x_j(t)$ to node i. Each node i updates its own value by forming a weighted average of its own value and the values it has just received from other nodes. We represent the sequence of communications between nodes by a sequence $G(t) = (\mathcal{N}, \mathcal{E}(t))$ of directed graphs, where $(j, i) \in \mathcal{E}(t)$ if and only if $a_{ij}(t) > 0$. Note that $(i, i) \in \mathcal{E}(t)$ for all t, and this condition will remain in effect throughout the paper.

Our next assumption requires that, following an arbitrary time t, and for any i, j, there is a sequence of communications through which node i will influence (directly or indirectly) the value held by node j.

ASSUMPTION 2.2 (connectivity). *For every $t \geq 0$, the graph $(\mathcal{N}, \cup_{s \geq t} \mathcal{E}(s)$ is strongly connected.*

Assumption 2.2 by itself is not sufficient to guarantee consensus (see Exercise 3.1, on page 517 of [3]). This motivates the following stronger version.

ASSUMPTION 2.3 (bounded interconnectivity times). *There is some B such that for all k, the graph $\big(\mathcal{N}, \mathcal{E}(kB) \cup \mathcal{E}(kB+1) \cup \cdots \cup \mathcal{E}((k+1)B-1)\big)$ is strongly connected.*

We note various special cases of possible interest.

Time-invariant model. In this model, introduced by DeGroot [10], the set of arcs $\mathcal{E}(t)$ is the same for all t; furthermore, the matrix $A(t)$ is the same for all t. In this case, we are dealing with the iteration $x := Ax$, where A is a stochastic matrix; in particular, $x(t) = A^t x(0)$. Under Assumptions 2.1 and 2.2, A is the transition probability matrix of an irreducible and aperiodic Markov chain. Thus, A^t converges to a matrix, all of whose rows are equal to the (positive) vector $\pi = (\pi_1, \ldots, \pi_n)$ of steady-state probabilities of the Markov chain. Accordingly, we have $\lim_{t\to\infty} x_i(t) = \sum_{i=1}^{n} \pi_i x_i(0)$.

Bidirectional model. In this case, we have $(i, j) \in \mathcal{E}(t)$ if and only if $(j, i) \in \mathcal{E}(t)$, and we say that the graph G is *symmetric*. Intuitively, whenever i communicates with j, there is a simultaneous communication from j to i.

Equal-neighbor model. Here,

$$a_{ij}(t) = \begin{cases} 1/d_i(t) & \text{if } j \in \mathcal{N}_i(t), \\ 0 & \text{if } j \notin \mathcal{N}_i(t), \end{cases}$$

where $\mathcal{N}_i(t) = \{j \mid (j, i) \in \mathcal{E}(t)\}$ is the set of nodes j (including i) whose value is taken into account by i at time t, and $d_i(t)$ is its cardinality. This model is a linear version of a model considered by Vicsek et al. [32]. Note that here the constant α of Assumption 2.1 can be take to be $1/n$.

THEOREM 2.4. *Under Assumptions* 2.1 *and* 2.3*, the agreement algorithm guarantees asymptotic consensus; that is, there exists some* c (*depending on* $x(0)$ *and on the sequence of graphs* $G(\cdot)$) *such that* $\lim_{t\to\infty} x_i(t) = c$ *for all* i.

Theorem 2.4 is presented in [31] and proved in [30], in a more general setting that allows for communication delays, under a slightly stronger version of Assumption 2.3; see also Chapter 7 of [3], and [31, 4], for extensions to the cases of communication delays and probabilistic dropping of packets. The above version of Assumption 2.3 was introduced in [18]. Under the additional assumption of a bidirectional model, the bounded interconnectivity time assumption is unnecessary, as established in [20, 6] for the bidirectional equal-neighbor model, and in [17, 25] for the general case.

3. Averaging with the agreement algorithm in fixed networks.

In this section, as well as in sections 4–8, we assume that the network topology is fixed, i.e., $G(t) = G$ for all t, and known. We consider the time-invariant version, $x := Ax$, of the agreement algorithm and discuss various ways that it can be used to solve the averaging problem. We show that an iteration $x := Ax$ that solves the consensus problem can be used in a simple manner to provide a solution to the averaging problem as well.

3.1. Using a doubly stochastic matrix.

As remarked in section 2, with the time-invariant agreement algorithm $x := Ax$, we have

$$\lim_{t\to\infty} x_i(t) = \sum_{i=1}^{n} \pi_i x_i(0) \qquad \forall\, i, \tag{3.1}$$

where π_i is the steady-state probability of node i in the Markov chain associated with the stochastic matrix A. It follows that we obtain a solution to the averaging problem if and only if $\pi_i = 1/n$ or every i. Since π is a left eigenvector of A, with eigenvalue equal to 1, this requirement translates into the property $\mathbf{1}^T A = \mathbf{1}^T$, where $\mathbf{1}$ is the vector with all components equal to 1. Equivalently, the matrix A needs to be doubly stochastic. A particular choice of a doubly stochastic matrix has been proposed in [27] (see also [8]); it is discussed further in sections 7 and 9.

3.2. The scaled agreement algorithm.

Suppose that the graph G is fixed a priori and that there is a system designer or other central authority who chooses a stochastic matrix A offline, computes the associated steady-state probability vector (assumed unique and positive), and disseminates the value of $n\pi_i$ to each node i.

Suppose next that the nodes execute the agreement algorithm $\overline{x} := A\overline{x}$, using the matrix A, but with the initial value $x_i(0)$ of each node i replaced by

$$\overline{x}_i(0) = \frac{x_i(0)}{n\pi_i}. \tag{3.2}$$

Then, the value $\overline{x}_i(t)$ of each node i converges to

$$\lim_{t\to\infty} \overline{x}_i(t) = \sum_{i=1}^{n} \pi_i \overline{x}_i(0) = \frac{1}{n}\sum_{i=1}^{n} x_i(0),$$

and we therefore have a valid averaging algorithm. This establishes that any (time-invariant) agreement algorithm for the consensus problem translates into an algorithm for the averaging problem as well. The following are two possible drawbacks of the scheme we have just described:

(a) If some of the $n\pi_i$ are very small, then some of the initial $\overline{x}_i(0)$ will be very large, which can lead to numerical difficulties [16].

(b) The algorithm requires some central coordination in order to choose A and compute π.

The algorithm provided in the next subsection provides a remedy for both of the above drawbacks.

3.3. Using two parallel passes of the agreement algorithm. Given a fixed graph G, let A be the matrix that corresponds to the time-invariant, equal-neighbor, bidirectional model (see section 2 for definitions); in particular, if $(i,j) \in \mathcal{E}$, then $(j,i) \in \mathcal{E}$, and $a_{ij} = 1/d_i$, where d_i is the cardinality of $\mathcal{N}_i$. Under Assumptions 2.1 and 2.2, the stochastic matrix A is irreducible and aperiodic (because $a_{ii} > 0$ for every i). Let $E = \sum_{i=1}^{n} d_i$. It is easily verified that the vector π with components $\pi_i = d_i/E$ satisfies $\pi^T = \pi^T A$ and is therefore equal to the vector of steady-state probabilities of the associated Markov chain.

The following averaging algorithm employs two parallel runs of the agreement algorithm, with different, but locally determined, initial values.

Algorithm 3.1.

(a) *Each node i sets $y_i(0) = 1/d_i$ and $z_i(0) = x_i(0)/d_i$.*

(b) *The nodes run the agreement algorithms $y(t+1) = Ay(t)$ and $z(t+1) = Az(t)$.*

(c) *Each node sets $x_i(t) = z_i(t)/y_i(t)$.*

We have

$$\lim_{t\to\infty} y_i(t) = \sum_{i=1}^{n} \pi_i y_i(0) = \sum_{i=1}^{n} \frac{d_i}{E}\cdot\frac{1}{d_i} = \frac{n}{E}$$

and

$$\lim_{t\to\infty} z_i(t) = \sum_{i=1}^{n} \pi_i z_i(0) = \sum_{i=1}^{n} \frac{d_i}{E}\cdot\frac{x_i(0)}{d_i} = \frac{1}{E}\sum_{i=1}^{n} x_i(0).$$

This implies that

$$\lim_{t\to\infty} x_i(t) = \frac{1}{n}\sum_{i=1}^{n} x_i(0);$$

i.e., we have a valid averaging algorithm. Note that the iteration $y := Ay$ need not be repeated if the network remains unchanged and the averaging algorithm is to be executed again with different initial opinions. Finally, if n and E are known by all nodes, the iteration $y := Ay$ is unnecessary, and we could just set $y_i(t) = n/E$.

4. Definition of the convergence rate and the convergence time. The convergence rate of any of the algorithms discussed in section 3 is determined by the convergence rate of the matrix powers A^t. In this section, we give a definition of the convergence rate (and convergence time) and provide a tool for bounding the convergence rate. As should be apparent from the discussion in section 3, there is no reason to restrict to doubly stochastic matrices, or even to nonnegative matrices. We therefore start by specifying the class of matrices that we will be interested in.

Consider a matrix A with the following property: For every $x(0)$, the sequence generated by letting $x(t+1) = Ax(t)$ converges to $c\mathbf{1}$ for some scalar c. Such a matrix corresponds to a legitimate agreement algorithm and can be employed in the scheme of section 3.2 to obtain an averaging algorithm, as long as 1 is an eigenvalue of A with multiplicity 1, and the corresponding left eigenvector, denoted by π, has nonzero entries. Because of the above assumed convergence property, all other eigenvalues must have magnitude less than 1. Note, however, that we allow A to have some negative entries.

Suppose that A has the above properties. Let $1 = \lambda_1, \lambda_2, \ldots, \lambda_n$, be the eigenvalues of A, sorted in order of decreasing magnitude. We also let X be the set of vectors of the form $c\mathbf{1}$, i.e., with equal components. Given such a matrix A, we define its *convergence rate*, ρ, by

$$\rho = \sup_{x(0)\notin X} \lim_{t\to\infty} \left(\frac{\|x(t) - x^*\|_2}{\|x(0) - x^*\|_2} \right)^{1/t}, \tag{4.1}$$

where x^* stands for $\lim_{t\to\infty} x(t)$. As is well known, we have $\rho = \max\{|\lambda_2|, |\lambda_n|\}$.

We also define the *convergence time*, $T_n(\epsilon)$, by

$$T_n(\epsilon) = \min \left\{ \tau : \frac{\|x(t) - x^*\|_\infty}{\|x(0) - x^*\|_\infty} \le \epsilon \quad \forall\, t \ge \tau,\ \forall\, x(0) \notin X \right\}.$$

Although we use the infinity norm to define the convergence time, bounds for other norms can be easily obtained from our subsequent results, by using the equivalence of norms.

Under the above assumptions, a result from [33] states

$$\rho = \max\{|\lambda_2|, |\lambda_n|\}.$$

To study ρ, therefore, we must develop techniques to bound the eigenvalues of the matrix A. To this end, we will be using the following result from [23]. We present here a slightly more general version and include a proof for completeness.

THEOREM 4.1. *Consider an $n \times n$ matrix A, and let $\lambda_1, \lambda_2, \ldots, \lambda_n$, be its eigenvalues, sorted in order of decreasing magnitude. Suppose that the following conditions hold:*

(a) *We have $\lambda_1 = 1$ and $A\mathbf{1} = \mathbf{1}$.*

(b) *There exists a positive vector π such that $\pi^T A = \pi^T$.*

(c) *For every i and j, we have $\pi_i a_{ij} = \pi_j a_{ji}$.*

Let

$$S = \left\{ x \;\middle|\; \sum_{i=1}^n \pi_i x_i = 0,\ \sum_{i=1}^n \pi_i x_i^2 = 1 \right\}.$$

Then, all eigenvalues of A are real, and

$$\lambda_2 = 1 - \frac{1}{2} \min_{x \in S} \sum_{i=1}^{n} \sum_{j=1}^{n} \pi_i a_{ij} (x_i - x_j)^2. \tag{4.2}$$

In particular, for any vector y that satisfies $\sum_{i=1}^{n} \pi_i y_i = 0$, we have

$$\lambda_2 \geq 1 - \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \pi_i a_{ij} (y_i - y_j)^2}{2 \sum_{i=1}^{n} \pi_i y_i^2}. \tag{4.3}$$

Proof. Let D be a diagonal matrix whose ith diagonal entry is π_i. Condition (c) yields $DA = A^T D$. We define the inner product $\langle \cdot, \cdot \rangle_\pi$ by $\langle x, y \rangle_\pi = x^T D y$. We then have

$$\langle x, Ay \rangle_\pi = x^T D A y = x^T A^T D y = \langle Ax, y \rangle_\pi.$$

Therefore, A is self-adjoint with respect to this inner product, which proves that A has real eigenvalues.

Since the largest eigenvalue is 1, with an eigenvector of $\mathbf{1}$, we use the variational characterization of the eigenvalues of a self-adjoint matrix (see Chapter 7, Theorem 4.3 of [29]) to obtain

$$\begin{aligned}\lambda_2 &= \max_{x \in S} \langle x, Ax \rangle_\pi \\ &= \max_{x \in S} \sum_{i=1}^{n} \sum_{j=1}^{n} \pi_i a_{ij} x_i x_j \\ &= \frac{1}{2} \max_{x \in S} \sum_{i=1} \sum_{j=1} \pi_i a_{ij} (x_i^2 + x_j^2 - (x_i - x_j)^2).\end{aligned}$$

For $x \in S$, we have

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \pi_i a_{ij} (x_i^2 + x_j^2) = 2 \sum_{i=1}^{n} \sum_{j=1}^{n} \pi_i a_{ij} x_i^2 = 2 \sum_{i=1}^{n} \pi_i x_i^2 = 2 \langle x, x \rangle_\pi = 2,$$

which yields

$$\lambda_2 = 1 - \frac{1}{2} \min_{x \in S} \sum_{i=1}^{n} \sum_{j=1}^{n} \pi_i a_{ij} (x_i - x_j)^2.$$

Finally, (4.3) follows from (4.2) by considering the vector

$$x_i = y_i \Big/ \sqrt{\left(\sum_{j=1}^{n} \pi_j y_j^2 \right)}. \qquad \square$$

Note that the bound of (4.3) does not change if we replace the vector y with αy for any $\alpha \neq 0$.

5. Convergence time for Algorithm 3.1. For the equal-neighbor, time-invariant, bidirectional model, tight bounds on the convergence rate were derived in [23].

Theorem 5.1 (see [23]). *Consider the equal-neighbor, time-invariant, bidirectional model on a connected graph with n nodes. The convergence rate satisfies*

$$\rho \leq 1 - \gamma_1 n^{-3},$$

where γ_1 is a constant independent of n. Moreover, there exists some $\gamma_2 > 0$ such that for every positive integer n, there exists an n-node connected symmetric graph for which

$$\rho \geq 1 - \gamma_2 n^{-3}.$$

Theorem 5.1 is proved in [23] for the case of symmetric graphs without self-arcs. It is not hard to check that essentially the same proof holds when self-arcs are present, with the only difference being in the values of the constants γ_1 and γ_2. This is intuitive because the effect of the self-arcs is essentially a "slowing down" of the Markov chain by a factor of at most 2, and therefore the convergence rate should stay the same.

Using some additional results on random walks, Theorem 5.1 leads to a tight bound (within a logarithmic factor) on the convergence time.

Corollary 5.2. *The convergence time for the equal-neighbor, time-invariant, symmetric model on a connected graph on n nodes satisfies*[2]

$$T_n(\epsilon) = O(n^3 \log(n/\epsilon)).$$

Furthermore, for every positive integer n, there exists an n-node connected graph for which

$$T_n(\epsilon) = \Omega(n^3 \log(1/\epsilon)).$$

Proof. The matrix A is the transition probability matrix for a random walk on the given graph, where, given the current state i, the next state is equally likely to be any of its neighbors (including i itself). Let $p_{ij}(t)$ be the (i,j)th entry of the matrix A^t. It is known that (see Theorem 5.1[3]of [21])

$$|p_{ij}(t) - \pi_j| \leq \sqrt{\frac{d_j}{d_i}} \rho^t. \tag{5.1}$$

Since $1 \leq d_i$ and $d_j \leq n$, we have

$$|p_{ij}(t) - \pi_j| \leq \sqrt{n} \rho^t$$

for all i, j, and t. Using the result of Theorem 5.1, we obtain

$$|p_{ij}(t) - \pi_j| \leq \sqrt{n}(1 - n^{-3})^t. \tag{5.2}$$

This implies that by taking $t = cn^3 \log(n/\epsilon)$, where c is a sufficiently large absolute constant, we will have $|p_{ij}(\tau) - \pi_j| \leq \epsilon/n$ for all i, j, and $\tau \geq t$.

[2] Throughout, log will stand for the base-2 logarithm.

[3] Theorem 5.1 of [21] is proved for symmetric graphs without self-arcs. However, the proof does not use the absence of self-arcs, and when they are present the same proof yields the same result. We refer the reader to the derivation of [21, section 3.1] for details.

Let $A^* = \lim_{t\to\infty} A^t$, and let $x^* = \lim_{t\to\infty} A^t x(0)$. Note that $A^*x(0) = x^* = A^t x^* = A^* x^*$ for all t. Then, with t chosen as above,

$$\begin{aligned}\|x(t) - x^*\|_\infty &= \|A^t(x(0) - x^*)\|_\infty \\ &= \|(A^t - A^*)(x(0) - x^*)\|_\infty \\ &\le \|A^t - A^*\|_1 \cdot \|x(0) - x^*\|_\infty \\ &\le \epsilon \|x(0) - x^*\|_\infty.\end{aligned}$$

This establishes the upper bound on $T_n(\epsilon)$.

For the lower bound, note that for every $(i, j) \in \mathcal{E}$, we have $\pi_i a_{ij} = (d_i/E)(1/d_i) = 1/E$, so that condition (c) in Theorem 5.1 is satisfied. It follows that A has real eigenvalues. Let $x(0)$ be a (real) eigenvector of A corresponding to the eigenvalue ρ. Then, $x(t) = A^t x(0) = \rho^t x(0)$, which converges to zero, i.e., $x^* = 0$. We then have

$$\frac{\|x(t) - x^*\|_\infty}{\|x(0) - x^*\|_\infty} = \rho^t.$$

By the second part of Theorem 5.1, there exists a graph for which $\rho \ge 1 - \gamma n^{-3}$, leading to the inequality $T_n(\epsilon) \ge cn^3 \log(1/\epsilon)$, for some absolute constant c. □

The $\Omega(n^3)$ convergence time of this algorithm is not particularly attractive. In the next section, we explore possible improvements in the convergence time by using different choices for the matrix A.

6. Convergence time for the scaled agreement algorithm. In this section, we consider the scaled agreement algorithm introduced in section 3.2. As in [33], we assume the presence of a system designer who chooses the matrix A so as to obtain a favorable convergence rate, subject to the condition $a_{ij} = 0$ whenever $(i, j) \notin \mathcal{E}$. The latter condition is meant to represent the network topology through which the nodes are allowed to communicate. Our goal is to characterize the best possible convergence rate guarantee. We will see that the convergence rate can be brought arbitrarily close to zero. However, if we impose a certain "numerical stability" requirement, the convergence time becomes $\Omega(n^2 \log(1/\epsilon))$ for a worst-case choice of the underlying graph. Furthermore, this worst-case lower bound applies even if we allow for matrices A in a much larger class than that considered in [33]. Finally, we will show that a convergence time of $O(n^2 \log(n/\epsilon))$ can be guaranteed in a simple manner, using a spanning tree.

6.1. Favorable but impractical convergence rates. In this section, we show that given a connected symmetric directed graph $G = (\mathcal{N}, \mathcal{E})$, there is an elementary way of choosing a stochastic matrix A for which ρ is arbitrarily close to zero.

We say that a directed graph is a *bidirectional spanning tree* if (a) it is symmetric, (b) it contains all self-arcs (i, i), and (c) we delete the self-arcs, ignore the orientation of the arcs, and remove duplicate arcs, in which case we are left with a spanning tree.

Without loss of generality, we assume that G is a bidirectional spanning tree; since G is symmetric and connected, this amounts to deleting some of its arcs, or, equivalently, setting $a_{ij} = 0$ for all deleted arcs (i, j).

Pick an arbitrary node, denoted by r, and designate it as the root. Consider an arc (i, j) and suppose that j lies on the path from i to the root. Let $\overline{a}_{ij} = 1$ and $\overline{a}_{ji} = 0$. Finally, let $\overline{a}_{rr} = 1$, and let $\overline{a}_{ii} = 0$ for $i \ne r$. This corresponds to a Markov chain in which the state moves deterministically towards the root. We have $\overline{A}^t = e_r \mathbf{1}^T$ for all $t \ge n$, where e_i is the ith basis vector. It follows that $\rho = 0$ and $T_n(\epsilon) \le n$.

However, this matrix $\overline{A}$ is not useful because the corresponding vector of steady-state probabilities has mostly zero entries, which prohibits the scaling discussed in section 3.2. Nevertheless, this is easily remedied by perturbing the matrix $\overline{A}$ as follows. For every $(i,j) \in \mathcal{E}$ with $i \neq j$ and $\overline{a}_{ij} = 0$, let $a_{ij} = \delta$, where δ is a small positive constant. For every i, there exists a unique j for which $\overline{a}_{ij} = 1$. For any such pair (i,j), we set $a_{ij} = 1 - \sum_{k=1}^{n} a_{ik}$ (which is nonnegative as long as δ is chosen small enough). We have thus constructed a new matrix A_δ which corresponds to a Markov chain whose transition diagram is a bidirectional spanning tree. Since the convergence rate ρ is an eigenvalue of the iteration matrix, and eigenvalues are continuous functions of matrix elements, we see that, for the matrix A_δ, the convergence rate ρ can be made as small as desired by choosing δ sufficiently small. Finally, since A_δ is a positive matrix, the corresponding vector of steady-state probabilities is positive.

To summarize, by choosing δ suitably small, we can choose a (stochastic) matrix A_δ with an arbitrarily favorable convergence rate, and which allows the application of the scaled agreement algorithm of section 3.2. It can be shown that the convergence time is linear in the following sense: For every ϵ, there exists some δ such that, for the matrix A_δ, the corresponding convergence time, denoted by $T_n(\epsilon;\delta)$, satisfies $T_n(\epsilon;\delta) \leq n$. Indeed, this is an easy consequence of the facts $\lim_{\delta \to 0}(A_\delta^n - \overline{A}^n) = 0$ and $T_n(\epsilon';0) \leq n$ for every $\epsilon' > 0$.[4]

However, note that as n gets larger, $n\pi_i$ may approach 0 at the nonroot nodes. The implementation of the scaling in (3.2) will involve division by a number which approaches 0, possibly leading to numerical difficulties. Thus, the resulting averaging algorithm may be undesirable. Setting averaging aside, the agreement algorithm based on A_δ, with δ small, is also undesirable; i.e., despite its favorable convergence rate, the final value on which consensus is reached is approximately equal to the initial value $x_r(0)$ of the root node. Such a "dictatorial" solution runs contrary to the motivation behind consensus algorithms.

6.2. A lower bound. In order to avoid the numerical issues raised above, we will now impose a condition on the dominant (and positive) left eigenvector π of the matrix A, and we require

$$n\pi_i \geq \frac{1}{C} \quad \forall\, i, \tag{6.1}$$

where C is a given constant with $C > 1$. This condition ensures that $n\pi_i$ does not approach 0 as n gets large, so that the initial conditions in the scaled agreement algorithm of section 3.2 are well behaved. Furthermore, in the context of consensus algorithms, condition (6.1) has an appealing interpretation: it requires that the initial value $x_i(0)$ of every node i have a nonnegligible impact on the final value $\lim_{t\to\infty} x_k(t)$, on which consensus is reached.[5]

We will now show that, under the additional condition (6.1), there are graphs for which the convergence time is $\Omega(n^2 \log(1/\epsilon))$. One may wonder whether a better convergence time is possible by allowing some of the entries of A to be negative. As

[4] Indeed, it is easy to see that by suitably choosing the root, we can make sure that convergence time is at most $\lceil d(G)/2 \rceil$ where $d(G)$ is the diameter of the graph G defined as the largest distance between any two vertices.

[5] In the case where A is the transition matrix of a reversible Markov chain, there is an additional interpretation. A reversible Markov chain may be viewed as a random walk on an undirected graph with edge-weights. Defining the degree of an vertex as the sum total of the weights incident upon it, the condition $n\pi_i \geq C$ is equivalent to requiring that each degree is lower bounded by a constant times the average degree.

the following result shows, negative entries do not help. The graph that we employ is a *line graph*, with arc set $\mathcal{E} = \{(i,j) \mid |i-j| \leq 1\}$.

THEOREM 6.1. *Consider an $n \times n$ matrix A such that $a_{ij} = 0$ whenever $|i-j| > 1$, and such that the graph with edge set $\{(i,j) \in \mathcal{E} \mid a_{ij} \neq 0\}$ is connected. Let $\lambda_1, \lambda_2, \ldots$ be its eigenvalues in order of decreasing modulus. Suppose that $\lambda_1 = 1$ and $A\mathbf{1} = 1$. Furthermore, suppose that there exists a vector π satisfying* (6.1) *such that $\pi^T A = \pi^T$. Then, there exist absolute constants c_1 and c_2 such that*

$$\rho \geq 1 - c_1 \frac{C}{n^2}$$

and

$$T_n(\epsilon) \geq c_2 \frac{n^2}{C} \log\left(\frac{1}{\epsilon}\right).$$

Proof. If the entries of A were all nonnegative, we would be dealing with a birth-death Markov chain. Such a chain is reversible, i.e., satisfies the detailed balance equations $\pi_i a_{ij} = \pi_j a_{ji}$ (condition (c) in Theorem 4.1). In fact the derivation of the detailed balance equations does not make use of nonnegativity; thus, detailed balance holds in our case as well.

Without loss of generality, we can assume that $\sum_{i=1}^n \pi_i = 1$. For $i = 1, \ldots, n$, let $y_i = i - \beta$, where β is chosen so that $\sum_{i=1}^n \pi_i y_i = 0$. We will make use of the inequality (4.3). Since $a_{ij} = 0$ whenever $|i-j| > 1$, we have

$$\sum_{i=1}^n \sum_{j=1}^n \pi_i a_{ij} (y_i - y_j)^2 \leq \sum_{i=1}^n \sum_{j=1}^n \pi_i a_{ij} = 1. \tag{6.2}$$

Furthermore,

$$\sum_{i=1}^n \pi_i y_i^2 \geq \frac{1}{nC} \sum_{i=1}^n y_i^2 = \frac{1}{nC} \sum_{i=1}^n (i-\beta)^2 \geq \frac{1}{nC} \sum_{i=1}^n \left(i - \frac{n+1}{2}\right)^2 \geq \frac{n^2}{12C}. \tag{6.3}$$

The next-to-last inequality above is an instance of the general inequality $\mathbf{E}[(X-\beta)^2] \geq \mathrm{var}(X)$ applied to a discrete uniform random variable X. The last inequality follows from the well-known fact $\mathrm{var}(X) = (n^2-1)/12$. Using the inequality (4.3) and (6.2)–(6.3), we obtain the desired bound on ρ.

For the bound on $T_n(\epsilon)$, we let $x(0)$ be a (real) eigenvector of A, associated with the eigenvalue λ_2, and proceed as in the end of the proof of Corollary 5.2. □

Remark. Note that if the matrix A is as in the previous theorem, it is possible for the iteration $x(t+1) = Ax(t)$ to not converge at all. Indeed, nothing in the argument precludes the possibility that the smallest eigenvalue is -1, for example. In such a case, the lower bounds of the theorem—derived based on bounding the second largest eigenvalue—still hold, as the convergence rate and time are infinite.

6.3. Convergence time for spanning trees. We finally show that an $O(n^2)$ convergence time guarantee is easily obtained by restricting to a spanning tree.

THEOREM 6.2. *Consider the equal-neighbor, time-invariant, bidirectional model on a bidirectional spanning tree. We have*

$$\rho \leq 1 - \frac{1}{3n^2}$$

and

$$T_n(\epsilon) = O\big(n^2 \log(n/\epsilon)\big).$$

Proof. In this context, we have $\pi_i = d_i/E$, where $E = \sum_{i=1}^n d_i = 2(n-1)+n < 3n$. (The factor 2 arises because we have arcs in both directions; the additional term n corresponds to the self-arcs.) As in the proof of Theorem 6.1, the detailed balance conditions $\pi a_{ij} = \pi_j a_{ji}$ hold, and we can apply Theorem 4.1. Note that (4.2) can be rewritten in the form

$$\lambda_2 = 1 - \frac{1}{2} \min_{\sum_i^n d_i x_i = 0, \sum_i^n d_i x_i^2 = 1} \sum_{(i,j)\in\mathcal{E}} (x_i - x_j)^2. \tag{6.4}$$

We use the methods of [23] to show that for trees, λ_2 can be upper bounded by $1 - 1/3n^2$. Indeed, suppose that x satisfies $\sum_i^n d_i x_i = 0$ and $\sum_i^n d_i x_i^2 = 1$, and let $x_{\max}$ be such that $|x_{\max}| = \max_i |x_i|$. Then,

$$1 = \sum_i d_i x_i^2 \le 3n x_{\max}^2,$$

and it follows that $|x_{\max}| \ge 1/\sqrt{3n}$. Without loss of generality, assume that $x_{\max} > 0$ (otherwise, replace each x_i by $-x_i$). Since $\sum_i d_i x_i = 0$, there exists some i for which $x_i < 0$; let us denote such a negative x_i by x_{neg}. Then,

$$\frac{1}{\sqrt{3n}} \le x_{\max} - x_{\text{neg}} = (x_{\max} - x_{k_1}) + (x_{k_1} - x_{k_2}) + \cdots + (x_{k_{r-1}} - x_{\text{neg}}), \tag{6.5}$$

where $k_1, k_2, \ldots, k_{r-1}$ are the nodes on the path from $x_{\max}$ to x_{neg}. By the Cauchy–Schwarz inequality,

$$\frac{1}{3n} \le \frac{n}{2} \sum_{(i,j)\in\mathcal{E}} (x_i - x_j)^2. \tag{6.6}$$

(The factor of 1/2 in the right-hand side arises because the sum includes both terms $(x_{k_i} - x_{k_{i+1}})^2$ and $(x_{k_{i+1}} - x_{k_i})^2$.) Thus,

$$\sum_{(i,j)\in\mathcal{E}} (x_i - x_j)^2 \ge \frac{2}{3n^2},$$

which proves the bound for the second largest eigenvalue.

For the smallest eigenvalue, recall that $a_{ii} \ge 1/n$ for every i. It follows that the matrix A is of the form $I/n + Q$, where I is the identity matrix and Q is a nonnegative matrix whose row sums are equal to $1 - 1/n$. Thus, all of the eigenvalues of Q have magnitude bounded above by $1 - 1/n$, which implies that the smallest eigenvalue of Q is bounded below by $-1 + 1/n$. We conclude that λ_n, the smallest eigenvalue of $I/n + Q$, satisfies

$$\lambda_n \ge -1 + \frac{2}{n} \ge -1 + \frac{2}{n^3}.$$

For the bound on the convergence time, we proceed as in the proof of Corollary 5.2. Let $p_{ij}(t)$ be the (i,j)th entry of A^t. Then,

$$|p_{ij}(t) - \pi_j| \le \sqrt{n}\left(1 - \frac{1}{3}n^{-2}\right)^t.$$

For a suitable absolute constant c and for $t \geq cn^2 \log(n/\epsilon)$, we obtain $|p_{ij}(t) - \pi(j)| \leq \epsilon/n$. The rest of the proof of Corollary 5.2 holds unchanged. $\square$

In light of the preceding theorem, we propose the following simple heuristic, with worst-case convergence time $O(n^2 \log(n/\epsilon))$, as an alternative to a more elaborate design of the matrix A.

ALGORITHM 6.3. We are given a symmetric graph G. We delete enough arcs to turn G into a bidirectional spanning tree, and then carry out the equal-neighbor, time-invariant, bidirectional consensus algorithm, with initial value $x_i(0)/n\pi_i$ at node i.

Let us remark that the $O(n^2 \log(n/\epsilon))$ bound (Theorem 6.2) on the convergence time of this heuristic is essentially tight (within a factor of $\log n$). Indeed, if the given graph is a line graph, then with our heuristic we have $n\pi_i = nd_i/E \geq 2/3$, and Theorem 6.1 provides an $\Omega(n^2 \log(1/\epsilon))$ lower bound.

7. Convergence time when using a doubly stochastic matrix.

We provide here a brief comparison of our methods with the following two methods that have been proposed in the literature and that rely on doubly stochastic matrices. Recall that doubly stochastic matrices give rise directly to an averaging algorithm, without the need for scaling the initial values.

(a) Reference [33] considers the case where the graph G is given and studies the problem of choosing a doubly stochastic matrix A for which the convergence rate ρ is smallest. In order to obtain a tractable (semidefinite programming) formulation, this reference imposes the further restriction that A be symmetric. For a doubly stochastic matrix, we have $\pi_i = 1/n$ for all i, so that condition (6.1) holds with $C = 1$. According to Theorem 6.1, there exists a sequence of graphs, for which we have $T_n(\epsilon) = \Omega(n^2 \log(1/\epsilon))$. We conclude that, despite the sophistication of this method, its worst-case guarantee is no better (ignoring the $\log n$ factor) than the simple heuristic we have proposed (Algorithm 6.3). On the other hand, for particular graphs, the design method of [33] may yield better convergence times.

(b) The following method was proposed in [27]. The nodes first agree on some value $\epsilon \in (0, 1/\max_i d_i)$. (This is easily accomplished in a distributed manner.) Then, the nodes iterate according to the equation

$$x_i(t+1) = (1 - \epsilon d_i)x_i(t) + \epsilon \sum_{j \in \mathcal{N}(i)\setminus\{i\}}^{n} x_j(t). \tag{7.1}$$

Assuming a connected graph, the iteration converges to consensus (this is a special case of Theorem 2.4). Furthermore, this iteration preserves the sum $\sum_{i=1}^{n} x_i(t)$. Equivalently, the corresponding matrix A is doubly stochastic, as required in order to have an averaging algorithm.

This algorithm has the disadvantage of uniformly small step sizes. If many of the nodes have degrees of the order of n, there is no significant theoretical difference between this approach and our Algorithm 3.1, as both have effective step sizes of order of $1/n$. On the other hand, if only a small number of nodes has large degree, then the algorithm in [27] will force *all* the nodes to take small steps. This drawback is avoided by our Algorithms 3.1 (section 3.3) and 6.3 (section 6.3). A comparison of the method of [27] with Algorithm 3.1 is carried out, through simulation experiments, in section 8.

8. Averaging with dynamic topologies.

In this section, we turn our attention to the more challenging case where communications are bidirectional but the network topology changes dynamically. Averaging algorithms for such a context have been considered previously in [24, 26].

As should be clear from the previous sections, consensus and averaging algorithms are intimately linked, with the agreement algorithm often providing a foundation for the development of an averaging algorithm. For this reason, we start by investigating the worst-case performance of the agreement algorithm in a dynamic environment. Unfortunately, as shown in section 8.1, its convergence time is not polynomially bounded, in general, even though it is an open question whether this is also the case when we restrict to symmetric graphs. Motivated by this negative result, we approach the averaging problem differently: we introduce an averaging algorithm based on "load balancing" ideas (section 8.2) and prove a polynomial bound on its convergence time (section 8.3).

8.1. Nonpolynomial convergence time for the agreement algorithm. We begin by formally defining the notion of "convergence time" for the agreement algorithm on dynamic graph sequences. Given a sequence of graphs $G(t)$ on n nodes such that Assumption 2.3 of section 2 is satisfied for some $B > 0$, and an initial condition $x(0)$, we define the convergence time $T_{G(\cdot)}(x(0), \epsilon)$ (for this particular graph sequence and initial condition) as the first time t when each node is within an ϵ-neighborhood of the final consensus, i.e., $\|x(t) - \lim_{t\to\infty} x(t)\|_\infty \leq \epsilon$. We then define the (worst-case) convergence time, $T_n(B, \epsilon)$, as the maximum value of $T_{G(\cdot)}(x(0), \epsilon)$, over all graph sequences $G(\cdot)$ on n nodes that satisfy Assumption 2.3 for that particular B, and over all initial conditions that satisfy $\|x(0)\|_\infty \leq 1$.

We focus our attention on the equal-neighbor version of the agreement algorithm. The result that follows shows that its convergence time is not bounded by a polynomial in n and B. In particular, if B is proportional to n, the convergence time increases faster than an exponential in n. We note that the upper bound in Theorem 8.1 is not a new result, but we include it for completeness, and for comparison with the lower bound, together with a proof sketch. Similar upper bounds have also been provided recently in [7], under slightly different assumptions on the graph sequence $G(\cdot)$.

THEOREM 8.1. *For the equal-neighbor agreement algorithm, there exist positive constants c_1 and c_2 such that for every n, B, and $1 > \epsilon > 0$,*

$$c_1 nB\Big(\frac{n-1}{2}\Big)^{B-1} \log\frac{1}{\epsilon} \leq T_n(B, \epsilon) \leq c_2 B n^{nB} \log\frac{1}{\epsilon}. \tag{8.1}$$

Proof. The upper bound follows by inspecting the proof of convergence of the agreement algorithm with the constant α of Assumption 2.1 set to $1/n$ (cf. [30, 4]).

We now prove the lower bound by exhibiting a sequence of graphs $G(t)$ and an initial vector $x(0)$, with $\|x(0)\|_\infty \leq 1$ for which $T_{G(\cdot)}(x(0), \epsilon) \geq c_1 nB(n/2)^{B-1} \log(1/\epsilon)$. We assume that n is even and $n \geq 4$. The initial condition $x(0)$ is defined as $x_i(0) = 1$ for $i = 1, \ldots, n/2$, and $x_i(0) = -1$ for $i = n/2 + 1, \ldots, n$.

(i) The graph $G(0)$, used for the first iteration, is shown in the left-hand side of Figure 8.1.

(ii) For $t = 1, \ldots, B-2$, we perform an equal-neighbor iteration, each time using the graph $G(t)$ shown in the right-hand side of Figure 8.1.

(iii) Finally, at time $B-1$, the graph $G(B-1)$ consists of the complete graph over the nodes $\{1, \ldots, n/2\}$ and the complete graph over the nodes $\{n/2 + 1, \ldots, n\}$.

(iv) This sequence of B graphs is then repeated, i.e., $G(t + kB) = G(t)$ for every positive integer k.

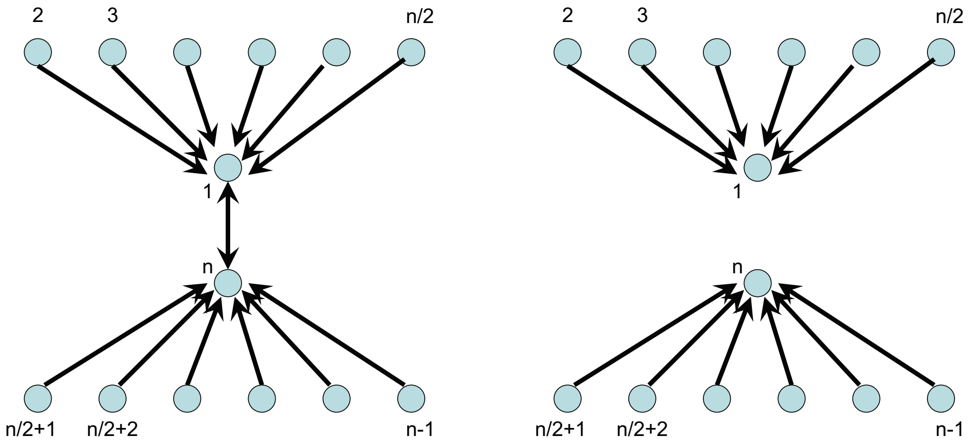


FIG. 8.1. *The diagram on the left is the graph $G(0)$. The diagram on the right is the graph $G(t)$ at times $t = 1, \ldots, B-2$. Self-arcs are not drawn but should be assumed present at every node.*

It is easily seen that this sequence of graphs satisfies Assumption 2.3, and that convergence to consensus is guaranteed.

At the end of the first iteration, we have $x_i(1) = x_i(0)$, for $i \neq 1, n$, and

$$x_1(1) = \frac{(n/2)-1}{(n/2)+1} = 1 - \frac{4}{n+2}, \qquad x_n(1) = -x_1(1). \tag{8.2}$$

Consider now the evolution of $x_1(t)$, for $t = 1, \ldots, B-2$, and let $\alpha(t) = 1 - x_1(t)$. We have

$$x_1(t+1) = \frac{1 \cdot (1-\alpha(t)) + (n/2-1) \cdot 1}{n/2} = 1 - (2/n)\alpha(t),$$

so that $\alpha(t+1) = 2\alpha(t)/n$. From (8.2), $\alpha(1) = 4/(n+2)$, which implies that $\alpha(B-1) = (2/n)^{B-2}$, or

$$x_1(B-1) = 1 - \frac{4}{n+2}\Big(\frac{2}{n}\Big)^{B-2}.$$

By symmetry,

$$x_n(B-1) = -1 + \frac{4}{n+2}\Big(\frac{2}{n}\Big)^{B-2}.$$

Finally, at time $B-1$, we iterate on the complete graph over nodes $\{1, \ldots, n/2\}$ and the complete graph over nodes $\{n/2+1, \ldots, n\}$. For $i = 2, \ldots, n/2$, we have $x_i(B-1) = 1$, and we obtain

$$x_i(B-1) = \frac{1 \cdot \Big(\dfrac{n}{2} - 1\Big) + 1 - \dfrac{4}{n+2}\Big(\dfrac{2}{n}\Big)^{B-2}}{n/2} = 1 - \frac{4}{n+2}\Big(\frac{2}{n}\Big)^{B-1}.$$

Similarly, for $i = (n/2)+1, \ldots, n$, we obtain

$$x_i(B-1) = -1 + \frac{4}{n+2}\Big(\frac{2}{n}\Big)^{B-2}.$$

Thus,

$$\frac{|\max_i x_i(B) - \min_i x_i(B)|}{|\max_i x_i(0) - \min_i x_i(0)|} = 1 - \frac{4}{n+2} \cdot \Big(\frac{2}{n}\Big)^{B-1}.$$

Moreover, because $x(B)$ is simply a scaled version of $x(0)$, it is clear that by repeating this sequence of graphs, we will have

$$\frac{|\max_i x_i(kB) - \min_i x_i(kB)|}{|\max_i x_i(0) - \min_i x_i(0)|} = \Big(1 - \frac{4}{n+2} \cdot \Big(\frac{2}{n}\Big)^{B-1}\Big)^k.$$

This readily implies that

$$T_{G(\cdot)(t)}(x(0), \epsilon) = \Omega\Big(nB\Big(\frac{n}{2}\Big)^{B-1} \log\frac{1}{\epsilon}\Big).$$

If n is odd, then $n' = n - 1$ is even. We apply the same initial condition and graph sequence as above to nodes $\{1, \dots, n'\}$. As for the additional node x_n, we let $x_n(0) = 0$ and make extra connections by connecting node n to nodes 1 and n' at time 0 with a bidirectional link. By repeating the analysis above, it can be verified that

$$T_{G(\cdot)(t)}(x(0), \epsilon) = \Omega\Big(nB\Big(\frac{n-1}{2}\Big)^{B-1} \log\frac{1}{\epsilon}\Big).$$

This concludes the proof. □

Both the upper and lower bounds in Theorem 8.1 display an exponential growth of the convergence time as a function of B. It is unclear, however, which of the two terms, n^B or n^{nB}, better captures the behavior of $T_n(B, \epsilon)$.

8.2. Polynomial time averaging in dynamic topologies. The algorithm we present here is a variation of an old *load balancing* algorithm (see [9] and Chapter 7.3 of [3]). Intuitively, a collection of processors with different initial loads tries to equalize its respective loads. As some of the highly loaded processors send some of their loads to their less loaded neighbors, the loads at different nodes tend to become equal. Similarly, at each step of our algorithm, each node offers some of its value to its neighbors and accepts or rejects such offers from its neighbors. Once an offer from i to j to send δ has been accepted, the updates $x_i := x_i - \delta$ and $x_j := x_j + \delta$ are executed.

We assume a time-varying sequence of graphs $G(t)$. We make only the following two assumptions on $G(\cdot)$: symmetry and bounded interconnectivity times (see section 2 for definitions). The symmetry assumption is natural if we consider, for example, communication between two nodes to be feasible whenever the nodes are within a certain distance of each other. The assumption of bounded interconnectivity times is necessary for an upper bound on the convergence time to exist (otherwise, we could insert infinitely many empty graphs $G(t)$, in which case convergence is arbitrarily slow for any algorithm).

We next describe formally the steps that each node carries out at each time t. For definiteness, we refer to the node executing the steps below as node A. Moreover, the instructions below sometimes refer to the "neighbors" of node A; this always means nodes other than A that are neighbors at time t, when the step is being executed (since $G(t)$ can change with t, the set of neighbors of A can also change). Let $\mathcal{N}_i(t) = \{j \neq i : (i,j) \in \mathcal{E}(t)\}$. Note that this is a little different from the definition of $\mathcal{N}_i(t)$ in earlier sections, in that i is no longer considered a neighbor of itself.

ALGORITHM 8.2. If $\mathcal{N}_A(t)$ is empty, node A does nothing at time t. Otherwise, node A carries out the following steps:

1. Node A broadcasts its current value x_A to all of its neighbors (every j with $j \in \mathcal{N}_A(t)$).
2. Node A finds a neighboring node B with the smallest value: $x_B = \min\{x_j : j \in \mathcal{N}_A(t)\}$. If $x_A \leq x_B$, then node A does nothing further at this step. If $x_B < x_A$, then node A makes an offer of $(x_A - x_B)/2$ to node B.
3. If node A does not receive any offers, it does nothing further at this step. Otherwise, it sends an acceptance to the sender of the largest offer and a rejection to all the other senders. It updates the value of x_A by adding the value of the accepted offer.
4. If an acceptance arrives for the offer made by node A, node A updates x_A by subtracting the value of the offer.

For concreteness, we use $x_i(t)$ to denote the value possessed by node i at the *beginning* of the above described steps. Accordingly, the value possessed by node i at the end of the above steps will be $x_i(t+1)$. The algorithm we have specified clearly keeps the value of $\sum_{i=1}^n x_i(t)$ constant. Furthermore, it is a valid averaging algorithm, as stated in Theorem 8.3 below. We do not provide a separate proof, because this result follows from the convergence time bounds in the next subsection.

THEOREM 8.3. *Suppose that each $G(t)$ is symmetric and that Assumption* 2.3 *(bounded interconnectivity times) holds. Then,* $\lim_{t\to\infty} x_i(t) = \frac{1}{n}\sum_{k=1}^n x_k(0)$ *for all i.*

8.3. Convergence time. We introduce the following "Lyapunov" function that quantifies the distance of the state $x(t)$ of the agents form the desired limit:

$$V(t) = \left\| x(t) - \frac{1}{n}\sum_{i=1}^n x_i(0)\mathbf{1} \right\|_2^2 .$$

Intuitively, $V(t)$ measures the variance of the values at the different nodes. Given a sequence of graphs $G(t)$ on n nodes, and an initial vector $x(0)$, we define the convergence time $T_{G(\cdot)}(x(0),\epsilon)$ as the first time t after which $V(\cdot)$ remains smaller than $\epsilon V(0)$:

$$T_{G(\cdot)}(x(0),\epsilon) = \min\big\{t \mid V(\tau) \leq \epsilon V(0)\ \forall\ \tau \geq t\big\}.$$

We then define the (worst-case) convergence time, $T_n(B,\epsilon)$, as the maximum value of $T_{G(\cdot)}(x(0),\epsilon)$ over all graph sequences $G(\cdot)$ on n nodes that satisfy Assumption 2.3 for that particular B, and over all initial conditions $x(0)$.

THEOREM 8.4. *There exists a constant $c > 0$ such that for every n and $1 > \epsilon > 0$, we have*

$$T_n(B,\epsilon) \leq cBn^3 \log\frac{1}{\epsilon}. \tag{8.3}$$

Proof. The proof is structured as follows. Without loss of generality, we assume that $\sum_{i=1}^n x_i(0) = 0$; this is possible because adding a constant to each x_i does not change the sizes of the offers or the acceptance decisions. We will show that $V(t)$ is nonincreasing in t, and that

$$V((k+1)B) \leq \Big(1 - \frac{1}{2n^3}\Big)V(kB) \tag{8.4}$$

for every nonnegative integer k. These two claims readily imply the desired result. To see this, note that if $V(t)$ decreases by a factor of $1-(1/2n^3)$ every B steps, then it decreases by a $\Theta(1)$ factor in Bn^3 steps. It follows that the time until $V(t)$ becomes less than $\epsilon V(0)$ is $O(Bn^3 \log(1/\epsilon))$. Finally, since $V(t)$ is nonincreasing, $V(t)$ stays below $\epsilon V(0)$ thereafter.

We first show that $V(t)$ is nonincreasing. We argue that while rejected offers clearly do not change $V(t)$, each accepted offer at time t results in a decrease of $V(t+1)$. While this would be straightforward to establish if there were a single accepted offer, a more complicated argument is needed to account for the possibility of multiple offers being simultaneously accepted. We will show that we can view the changes at time t as a result of a series of sequentially accepted offers, each of which results in a smaller value of V.

Let us focus on a particular time t. We order the nodes from smallest to largest, so that $x_1(t) \leq x_2(t) \leq \cdots \leq x_n(t)$, breaking ties arbitrarily. Let $A_i(t)$ be the size of the offer accepted by node i at time t (if any). If the node accepted no offers at time t, set $A_i(t)=0$. Furthermore, if $A_i(t)>0$, let $\mathcal{A}_i(t)$ be the index of the node whose offer node i accepted.

Let us now break time t into n *periods*. The ith period involves the updates caused by node i accepting an offer from node $\mathcal{A}_i(t)$. In particular, node i performs the update $x_i(t) := x_i(t)+A_i(t)$ and node $\mathcal{A}_i(t)$ performs the update $x_{\mathcal{A}_i(t)}(t) := x_{\mathcal{A}_i(t)}(t)-A_i(t)$.

We note that every offer accepted at time t appears in some period in the above sequence. We next argue that each offer decreases V. This will complete the proof that $V(t)$ is nonincreasing in t.

Let us suppose that in the ith period, node i accepts an offer from node $\mathcal{A}_i(t)$, which for simplicity we will denote by j. Because nodes only send offers to lower valued nodes, the inequality $x_j > x_i$ must hold at the beginning of time t, before time period 1. We claim that this inequality continues to hold when the ith time period is reached. Indeed, x_j is unchanged during periods $1,\ldots,i-1$ (it can only send one offer, which was to x_i; and if it receives any offers, their effects will occur in period j, which is after period i). Moreover, while the value of x_i may have changed in periods $1,\ldots,i-1$, it cannot have increased (since i is not allowed to accept more than one offer at any given time t). Therefore, the inequality $x_j > x_i$ still holds at the beginning of the ith period.

During the ith period, a certain positive amount is transferred from node j to node i. Since the transfer takes place from a higher-valued node to a lower-valued one, it is easily checked that the value of $x_i^2 + x_j^2$ (which is the contribution of these two nodes to V) is reduced. To summarize, we have shown that we can serialize the offers accepted at time t, in such a way that each accepted offer causes a reduction in V. It follows that $V(t)$ is nonincreasing.

We will now argue that at some time t in the interval $0,1,\ldots,B-1$, there will be some update (acceptance of an offer) that reduces $V(t)$ by at least $1/(2n^3)V(0)$. Without loss of generality, we assume $\max_i |x_i(0)| = 1$, so that all the values lie in the interval $[-1,+1]$. It follows that $V(0) \leq n$.

Since $\sum_{i=1}^n x_i(0) = 0$, it follows that $\min_i x_i(0) \leq 0$. Hence, the largest gap between any two consecutive $x_i(0)$ must be at least $1/n$. Thus, there exist some numbers a and b, with $b-a \geq 1/n$, and the set of nodes can be partitioned into two disjoint subsets S^- and S_+ such that $x_i(0) \leq a$ for all $i \in S_-$, and $x_i(0) \geq b$ for all $i \in S_+$. By Assumption 2.3, the graph with arcs $\bigcup_{s=0,\ldots,B-1} \mathcal{E}(s)$ is connected. Thus, there exists a first time $\tau \in \{0,1,\ldots,B-1\}$ at which there is a communication between some node $i \in S_-$ and some node $j \in S_+$, resulting in an offer from j to i. Up until that time, nodes in S_- have not interacted with nodes in S_+. It follows that $x_k(\tau) \leq a$ for all $k \in S_-$, and $x_k(\tau) \geq b$ for all $k \in S_+$. In particular, $x_i(\tau) \leq a$ and $x_j(\tau) \geq b$. There are two possibilities: either i accepts the offer from j, or i accepts some higher offer from some other node in S_+. In either case, we conclude that there is a first time $\tau \leq B - 1$, at which a node in S_- accepts an offer from a node in S_+.

Let us use plain x_i and x_j for the values at nodes i and j, respectively, at the beginning of period i of time τ. At the end of that period, the value at both nodes is equal to $(x_i + x_j)/2$. Thus, the Lyapunov function V decreases by

$$x_i^2 + x_j^2 - 2\Big(\frac{x_i + x_j}{2}\Big)^2 = \frac{1}{2}(x_i - x_j)^2 \geq \frac{1}{2}(b - a)^2 \geq \frac{1}{2n^2}.$$

At every other time and period, V is nonincreasing, as shown earlier. Thus, using the inequality $V(0) \leq n$,

$$V(B) \leq V(0) - \frac{1}{2n^2} \leq V(0)\Big(1 - \frac{1}{2n^3}\Big).$$

By repeating this argument over the interval $kB, \ldots, (k+1)B$, instead of the interval $0, \ldots, B$, we establish (8.4), which concludes the proof. □

9. Simulations.

We have proposed several new algorithms for the distributed consensus and averaging problems. For one of them, namely the spanning tree heuristic of section 6.3 (Algorithm 6.3), the theoretical performance has been characterized completely—see Theorem 6.2 and the discussion at the end of section 6.3. In this section, we provide simulation results for the remaining two algorithms.

9.1. Averaging in fixed networks with two passes of the agreement algorithm.

In section 3.3, we proposed a method for averaging in fixed graphs, based on two parallel executions of the agreement algorithm (Algorithm 3.1). We speculated in section 7 that the presence of a small number of high degree nodes would make the performance of our algorithm attractive relative to the algorithm of [27], which uses a step size proportional to the inverse of the largest degree. (Our implementation used a step size of $\epsilon = 1/2d_{\max}$.) Figure 9.1 presents simulation results for the two algorithms.

In each simulation, we first generate geometric random graph $G(n, r)$ by placing nodes randomly in $[0, 1]^2$ and connecting two nodes if they are at most r apart. We choose $r = \Theta(\sqrt{\log n/n})$, which is a standard choice for modeling wireless networks (cf. [11]).

We then change the random graph $G(n, r)$ by choosing n_d nodes at random ($n_d =$ 10 in both parts of Figure 9.1) and adding edges randomly to make the degree of these nodes linear in n; this is done by randomly inserting all possible edges incident to at least one node in n_d; each such edge in inserted independently with probability 1/3. We run the algorithm, with random starting values, uniformly distributed in $[0, 1]$, until the largest deviation from the mean is at most $\epsilon = 10^{-3}$.

Each outcome recorded in Figure 9.1 (for different values of n) is the average of three runs. We conclude that for such graphs, the convergence time of the algorithm in [27] grows considerably faster than the one proposed in this paper.

9.2. Averaging in time-varying random graphs.

We report here on simulations involving the load-balancing algorithm (Algorithm 8.2) on time-varying random

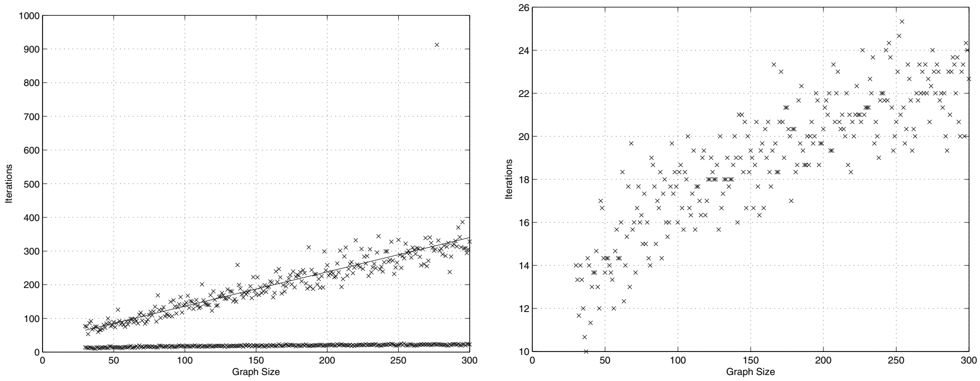


FIG. 9.1. *On the left: Comparison of averaging algorithms on a geometric random graph. The top line corresponds to the algorithm of* [27]*, and the bottom line (close to the horizontal axis) corresponds to using two parallel passes of the agreement algorithm (Algorithm* 3.1*). On the right: A blow-up of the performance of the agreement algorithm.*

graphs. In contrast to our previous simulations on static geometric graph, we test two time-varying models which simulate movement.

In both models, we select our initial vector $x(0)$ by choosing each component independently as a uniform random variable over $[0, 1]$. In our first model, at each time t, we independently generate an Erdös–Renyi random graph $G(t) = G(c, n)$ with $c = 3/4$. In the second model, at each time step we independently generate a geometric random graph with $G(n, r)$ with $r = \sqrt{\log n/n}$. In both models, if the largest deviation from the mean is at most $\epsilon = 10^{-3}$, we stop; otherwise, we perform another iteration of the load-balancing algorithm.

The results are summarized in Figure 9.2, where again each point represents the average of three runs. We conclude that in these random models, only a sublinear number of iterations appears to be needed.

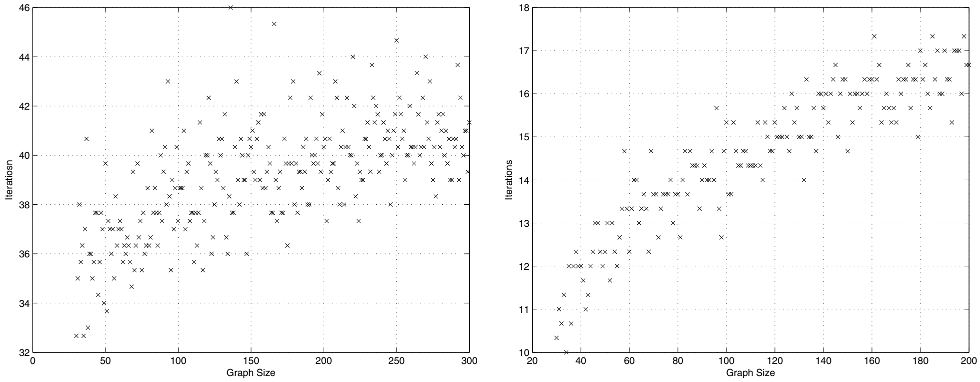


FIG. 9.2. *On the left: Averaging in time-varying Erdös–Renyi random graphs with the load balancing algorithm. Here* $c = 3/4$ *at each time* t*. On the right: Averaging in time-varying geometric random graphs with the load balancing algorithm. Here* $r = \sqrt{\log n/n}$*.*

10. Concluding remarks. In this paper we have considered a variety of consensus and averaging algorithms and studied their convergence rates. While our discussion was focused on averaging algorithms, several of our results pertain to the closely related consensus problem.

For the case of a fixed topology, we showed that averaging algorithms are easy to construct by using two parallel executions of the agreement algorithm for the consensus problem. We also saw that a reasonable performance guarantee can be obtained by using the equal-neighbor agreement algorithm on a spanning tree, as opposed to a more sophisticated design.

For the case of a fixed topology, the choice of different algorithms is not a purely mathematical issue; one must also take into account the extent to which one is able to design the algorithm offline and provide suitable instructions to each node. After all, if the nodes are able to set up a spanning tree, there are simple distributed algorithms, involving two sweeps along the tree, in opposite directions, with which the sum of their initial values can be computed and disseminated [3], thus eliminating the need for an iterative algorithm. On the other hand, in less structured environments, with the possibility of occasional changes in the system topology, iterative algorithms can be more resilient. For example, the equal-neighbor agreement algorithm adjusts itself naturally when the topology changes.

In the face of a changing topology (possibly at each time step), the agreement algorithm continues to work properly, under minimal assumptions (see Theorem 2.4). On the other hand, its worst-case convergence time may suffer severely (cf. section 8.1). Furthermore, it is not apparent how to modify the agreement algorithm and obtain an averaging algorithm without sacrificing linearity and/or allowing some additional memory at the nodes. In section 8, we introduced an averaging algorithm, which is nonlinear but leads to a rather favorable (and, in particular, polynomial) convergence time bound. In view of the favorable performance observed in our simulation results, it would also be interesting to characterize the average performance of this algorithm, under a probabilistic mechanism for generating the graphs $G(t)$, similar to the one in our simulations.

Something to notice about Algorithm 8.2 is that it requires the topology to remain fixed during the exchange of offers and acceptances/rejections that happens at each step. On the other hand, without such an assumption, or without introducing a much larger memory at each node (which would allow for flooding of individual values), an averaging algorithm may well turn out to be impossible.

REFERENCES

[1] P.-A. Bliman and D. Angeli, *Convergence Speed of Unsteady Distributed Consensus: Decay Estimate Along the Settling Spanning-Trees*, http://arxiv.org/abs/math.OC/0610854 (2007).

[2] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, *Randomized gossip algorithms*, IEEE Trans. Inform. Theory, 52 (2006), pp. 2508–2530.

[3] D. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[4] V. D. Blondel, J. M. Hendrickx, A. Olshevsky, and J. N. Tsitsiklis, *Convergence in multiagent coordination, consensus, and flocking*, in Proceedings of the Joint 44th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC'05), Seville, Spain, 2005, pp. 2996–3000.

[5] J. Cortes, *Finite-time convergent gradient flows with applications to network consensus*, Automatica, 42 (2006), pp. 1993–2000.

[6] M. Cao, A. S. Morse, and B. D. O. Anderson, *Coordination of an asynchronous, multi-agent system via averaging*, in Proceedings of the 16th International Federation of Automatic Control World Congress (IFAC), Prague, Czech Republic, 2005.

[7] M. Cao, D. A. Spielman, and A. S. Morse, *A lower bound on convergence of a distributed network consensus algorithm*, in Proceedings of the Joint 44th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC'05), Seville, Spain, 2005, pp. 2356–2361.

[8] C. Gao, J. Cortés, and F. Bullo, *Notes on averaging over acyclic digraphs and discrete coverage control*, Automatica, 44 (2008), pp. 2120–2127.

[9] G. Cybenko, *Dynamic load balancing for distributed memory multiprocessors*, J. Parallel Distribut. Comput., 7 (1989), pp. 279–301.

[10] M. H. DeGroot, *Reaching a Consensus*, J. Amer. Statist. Assoc., 69 (1974), pp. 118–121.

[11] A. G. Dimakis, A. D. Sarwate, and M. J. Wainwright, *Geographic gossip: Efficient aggregation for sensor networks*, in Proceedings of the Fifth International Conference on Information Processing in Sensor Networks (IPSN), Nashville, TN, 2006, pp. 69–76.

[12] P. Erdös and A. Renyi, *On the evolution of random graphs*, Magyar Tud. Akad. Mat. Kutató Int. Közl., 5 (1960), pp. 17–61.

[13] L. Fang and P. Antsaklis, *On communication requirements for multi-agent consensus seeking*, in Proceedings of the Workshop on Networked Embedded Sensing and Control, Lecture Notes in Control and Inform. Sci. 331, Springer, Berlin, 2006, pp. 53–67.

[14] A. Ganguli, S. Susca, S. Martinez, F. Bullo, and J. Cortes, *On collective motion in sensor networks: Sample problems and distributed algorithms*, in Proceedings of the Joint 44th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC'05), Seville, Spain, 2005, pp. 4239–4244.

[15] D. Grünbaum, S. Viscido, and J. K. Parrish, *Extracting interactive control algorithms from group dynamics of schooling fish*, in Cooperative Control, V. Kumar, N. Leonard, and A. S. Morse, eds., Lecture Notes in Control and Inform. Sci. 309, Springer, New York, 2005, pp. 103–117.

[16] N. Higham, *Accuracy and Stability of Numerical Algorithms*, 2nd ed., SIAM, Philadelphia, 2002.

[17] J. M. Hendrickx and V. D. Blondel, *Convergence of linear and non-linear versions of Vicsek model*, in Proceedings of 17th International Symposium on Mathematical Theory of Networks and Systems (MTNS'06), Kyoto, Japan, pp. 1229–1240. Available online at http://www-ics.acs.i.kyoto-u.ac.jp/mtns06/papers/0156.pdf.

[18] A. Jadbabaie, J. Lin, and A. S. Morse, *Coordination of groups of mobile autonomous agents using nearest neighbor rules*, IEEE Trans. Automat. Control, 48 (2003), pp. 988–1001.

[19] Y. Kim and M. Mesbahi, *On maximizing the second smallest eigenvalue of a state-dependent graph Laplacian*, IEEE Trans. Automat. Control, 51 (2006), pp. 116–120.

[20] S. Li and H. Wang, *Multi-Agent Coordination Using Nearest-Neighbor Rules: Revisiting the Vicsek Model*, http://arxiv.org/abs/cs.MA/0407021 (2004).

[21] L. Lovász, *Random walks on graphs: A survey*, in Combinatorics, Paul Erdös is Eighty, Vol. 2, D. Miklós, V. T. Sós, and T. Szõnyi, eds., János Bolyai Mathematical Society, Budapest, 1996, pp. 353–398.

[22] Y. Liu and Y. R. Yang, *Reputation propagation and agreement in mobile ad hoc networks*, in Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC), 2003, pp. 1510–1515.

[23] H. J. Landau and A. M. Odlyzko, *Bounds for eigenvalues of certain stochastic matrices*, Linear Algebra Appl., 38 (1981), pp. 5–15.

[24] M. Mehyar, D. Spanos, J. Pongsajapan, S. H. Low, and R. M. Murray, *Distributed averaging on asynchronous communication networks*, in Proceedings of the Joint 44th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC'05), Seville, Spain, 2005, pp. 7446–7451.

[25] L. Moreau, *Consensus seeking in multi-agent systems using dynamically changing interaction topologies*, IEEE Trans. Automat. Control, 50 (2005), pp. 169–182.

[26] C. Moallemi and B. Van Roy, *Consensus propagation*, IEEE Trans. Inform. Theory, 52 (2006), pp. 4753–4766.

[27] R. Olfati-Saber and R. M. Murray, *Consensus problems in networks of agents with switching topology and time-delays*, IEEE Trans. Automat. Control, 49 (2004), pp. 1520–1533.

[28] R. Olfati-Saber, J. A. Fax, and R. M. Murray, *Consensus and cooperation in networked multi-agent systems*, in Proc. IEEE, 95 (2007), pp. 215–233.

[29] S. Treil, *Linear Algebra Done Wrong*, http://www.math.brown.edu/$\sim$treil/papers/LADW/LADW.html.

[30] J. N. Tsitsiklis, *Problems in Decentralized Decision Making and Computation*, Ph.D. Thesis, Department of EECS, MIT, Cambridge, MA, 1984.

[31] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans, *Distributed asynchronous deterministic and stochastic gradient optimization algorithms*, IEEE Trans. Automat. Control, 31 (1986), pp. 803–812.

[32] T. Vicsek, E. Czirok, E. Ben-Jacob, I. Cohen, and O. Shochet, *Novel type of phase transitions in a system of self-driven particles*, Phys. Rev. Lett., 75 (1995), pp. 1226–1229.

[33] L. Xiao and S. Boyd, *Fast linear iterations for distributed averaging*, Systems Control Lett., 53 (2004), pp. 65–78.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 1, pp. 56–76

© 2009 Society for Industrial and Applied Mathematics

CONSENSUS OPTIMIZATION ON MANIFOLDS*

ALAIN SARLETTE† AND RODOLPHE SEPULCHRE†

Abstract. The present paper considers distributed consensus algorithms that involve N agents evolving on a connected compact homogeneous manifold. The agents track no external reference and communicate their relative state according to a communication graph. The consensus problem is formulated in terms of the extrema of a cost function. This leads to efficient gradient algorithms to synchronize (i.e., maximizing the consensus) or balance (i.e., minimizing the consensus) the agents; a convenient adaptation of the gradient algorithms is used when the communication graph is directed and time-varying. The cost function is linked to a specific centroid definition on manifolds, introduced here as the *induced arithmetic mean*, that is easily computable in closed form and may be of independent interest for a number of manifolds. The special orthogonal group $SO(n)$ and the Grassmann manifold $Grass(p,n)$ are treated as original examples. A link is also drawn with the many existing results on the circle.

Key words. consensus algorithms, decentralized control, swarm control, synchronization, differential geometry, mean on manifolds, special orthogonal group, Grassmann manifold

AMS subject classifications. 93A14, 68W15, 49Q99, 43A07, 93B27, 14M15

DOI. 10.1137/060673400

1. Introduction.

The distributed computation of means/averages of datasets (in an algorithmic setting) and the *synchronization* of a set of agents (in a control setting)—i.e., driving all the agents to a common point in state space—are ubiquitous tasks in current engineering problems. Likewise, spreading a set of agents in the available state space—linked to the definition of *balancing* in section 4—is a classical problem of growing interest. Practical applications include autonomous swarm/formation operation (see, e.g., [42, 28, 22, 23, 25]), distributed decision making (see, e.g., [35, 47]), neural and communication networks (see, e.g., [46, 19]), clustering and other reduction methods (see, e.g., [17]), optimal covering or coding (see, e.g., [3, 4, 11, 12]), and other fields where averaging/synchronizing or distributing a set of points appears as a subproblem. In a modeling framework, the understanding of synchronization, or more generally, swarm behavior, has also led to many important studies (see, e.g., [26, 45, 48]).

Synchronization algorithms are well understood in Euclidean spaces (see, e.g., [33, 32, 47, 35]). They are based on the natural definition and distributed computation of the centroid in $\mathbb{R}^m$. However, many of the applications above involve manifolds that are not homeomorphic to an Euclidean space. Even for formations moving in $\mathbb{R}^2$ or $\mathbb{R}^3$, the agents' orientations evolve in a manifold $SO(2) \cong S^1$ or $SO(3)$. Balancing only makes sense on compact state spaces; though many theoretical results concern convex or star-shaped subsets of $\mathbb{R}^m$ (see, e.g., [12]), most applications involve compact

*Received by the editors October 26, 2006; accepted for publication (in revised form) April 3, 2008; published electronically February 11, 2009. This paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The scientific responsibility rests with its authors.

http://www.siam.org/journals/sicon/48-1/67340.html

†Department of Electrical Engineering and Computer Science (Montefiore Institute), University of Liège, Sart-Tilman Bldg. B28, B-4000 Liège, Belgium (alain.sarlette@ulg.ac.be, r.sepulchre@ulg.ac.be). The first author is supported as an FNRS fellow (Belgian Fund for Scientific Research).

manifolds. It seems that the study of global synchronization or balancing in non-Euclidean manifolds is not widely covered in the literature, except for the circle.

The present paper proposes algorithms for global synchronization and balancing—grouped under the term *consensus*—on connected compact homogeneous manifolds. A homogeneous manifold $\mathcal{M}$ is isomorphic to the quotient of two Lie groups. Intuitively, it is a manifold on which "all points are equivalent." This makes the problem symmetric with respect to the absolute position on the manifold and allows us to focus on *configurations* of the swarm, i.e., *relative positions* of the agents.

The main idea is to embed $\mathcal{M}$ in $\mathbb{R}^m$ and measure distances between agents in $\mathbb{R}^m$ in order to build a convenient cost function for an optimization-based approach. The related centroid on $\mathcal{M}$ may be interesting on its own; it is therefore studied in more detail in section 3 of the paper.

Throughout the paper, the abstract concepts are illustrated on the special orthogonal group $SO(n)$, the Grassmann manifold $Grass(p, n)$ of p-dimensional vector spaces in $\mathbb{R}^n$, and sometimes the circle S^1, which is in fact isomorphic to both $SO(2)$ and $Grass(1, 2)$. Other manifolds to which the present framework could be applied include the n-dimensional spheres S^n and the connected compact Lie groups. The circle S^1 is the simplest example; it links the present work to existing results in [42, 41, 39, 43]. $SO(n)$ is important in control applications as the natural state space for orientations of n-dimensional rigid bodies. $Grass(p, n)$ appears in algorithmic problems; [11] mentions the optimal placement of N laser beams for cancer treatment and the projection of multidimensional data on N representative planes as practical applications of optimal distributions on $Grass(p, n)$.

The paper is organized as follows. Previous work is briefly reviewed in section 1.1. Section 2 introduces concepts and notations about graph theory, $SO(n)$, and $Grass(p, n)$. Section 3 is devoted to the induced arithmetic mean. A definition of consensus is presented in section 4. Section 5 introduces a cost function to express the consensus problem in an optimization setting. Section 6 derives gradient algorithms based on this cost function, with the only communicated information being the relative positions of interconnected agents; convergence is proved for any connected, fixed, and undirected communication graph. Algorithms whose convergence properties can be guaranteed under possibly directed, time-varying, and disconnected communication graphs are presented in section 7; they employ an auxiliary variable that evolves in the embedding space $\mathbb{R}^m$.

1.1. Previous work. Most of the work related to synchronization and balancing on manifolds concerns the circle S^1. The most extensive literature on the subject derives from the Kuramoto model (see [44] for a review). Recently, however, synchronization on the circle has been considered from a control perspective, with the state variables representing headings of agents in the plane. Most results cover *local* convergence [22, 33]. An interesting set of *globally* convergent algorithms in $SE(2) = S^1 \times \mathbb{R}^2$ is presented in [42], but they require all-to-all communication. Some problems related to global discrete-time synchronization on S^1 under different communication constraints are discussed in [37], where connections of the control problem with various existing models are made. Stronger results are presented in [39] for global synchronization and balancing on S^1 with varying, directed communication links, at the cost of introducing estimator variables which communicating agents must exchange. Finally, [43] presents results on $SE(2)$ similar to those of [42] but under relaxed communication assumptions and using, among others, the estimator strategy of [39, 40].

Several authors have already presented algorithms that asymptotically synchronize satellite attitudes, involving the rotation group $SO(3)$. They often rely on tracking a common external reference (see, e.g., [27]) or leader (see, e.g., [5, 25]). The use of the convenient, but nonunique, quaternion representation for $SO(3)$ produces unwanted artifacts in the satellites' motions. Attitude synchronization without common references and quaternion artifacts is studied in [34]; using the same distance measure as the present work, an artificial coupling potential is built to establish local stability. All these approaches explicitly incorporate the second-order rigid-body dynamics. In accordance with the consensus approach, the present paper reduces the agents to first-order kinematic models to focus on (almost) global convergence for various agent interconnections, without any leader or external reference. Application of this framework in a mechanical setting is discussed in a separate paper [38].

Synchronization or balancing on a manifold $\mathcal{M}$ is closely related to the definition and computation of a mean or centroid of points on $\mathcal{M}$, a basic problem that has attracted somewhat more attention, as can be seen from [20, 9, 15], among others.

A key element of the present paper is the computation of a centroid in the embedding space $\mathbb{R}^m$ of $\mathcal{M}$, which is then projected onto $\mathcal{M}$. This is connected to the "projected arithmetic mean" defined in [31] for $SO(3)$. The idea of computing of statistics in a larger and simpler embedding manifold (usually Euclidean space) and projecting the result back onto the original manifold goes back to 1972 [13].

A short example in [1] addresses the computation of a "centroid of subspaces" without much theoretical analysis; in fact, our algorithms on $Grass(p, n)$ are similar and can eventually be viewed as generalizing the developments in [1] in the framework of consensus and synchronization. More recently, [17] used the centroid associated to the projector representation of $Grass(p, n)$, exactly as is done below but without going into theoretical details, to compute the cluster centers in a clustering algorithm. The distance measure associated with this centroid on $Grass(p, n)$ is called the *chordal distance* in [11, 4], where it is used to derive optimal distributions ("packings") of N agents on some specific Grassmann manifolds.

Finally, the topic of optimization-based algorithm design on manifolds has considerably developed over the last decades (see, e.g., [6, 14] and the books [18, 2]).

2. Preliminaries.

2.1. Elements of graph theory.

Consensus among a group of agents depends on the available communication links. When considering limited agent interconnections, it is customary to represent communication links by means of a *graph.* The graph G is composed of N *vertices* (the N agents) and contains the *edge* (j, k) if agent j sends information to agent k, which is denoted $j \rightsquigarrow k$. A positive *weight* a_{jk} is associated with each edge (j, k) to obtain a weighted graph; the weight is extended to any pair of vertices by imposing $a_{jk} = 0$ iff (j, k) does not belong to the edges of G. The full notation for the resulting *digraph* (directed graph) is $G(V, E, A)$, where $V = \{\text{vertices}\}$, $E = \{\text{edges}\}$, and A, containing the a_{jk}, is the *adjacency matrix.* The convention $a_{kk} = 0 \ \forall k$ is assumed for the representation of communication links.

The *out-degree* of a vertex k is the quantity $d_k^{(o)} = \sum_{j=1}^N a_{kj}$ of information leaving k towards other agents; its *in-degree* is the quantity $d_k^{(i)} = \sum_{j=1}^N a_{jk}$ of information received by k from other agents. These degrees can be assembled in diagonal matrices $D^{(o)}$ and $D^{(i)}$. A graph is *balanced* if $D^{(o)} = D^{(i)}$. This is satisfied in particular by *undirected* graphs, for which $A = A^T$. A graph is *bidirectional* if $(j, k) \in E \Leftrightarrow (k, j) \in E$ (but not necessarily $A = A^T$).

The *Laplacian* L of a graph is $L = D - A$. For directed graphs, $D^{(i)}$ or $D^{(o)}$ can be used, leading to the in-Laplacian $L^{(i)} = D^{(i)} - A$ and the out-Laplacian $L^{(o)} = D^{(o)} - A$. By construction, $L^{(i)}$ has zero column sums and $L^{(o)}$ has zero row sums. The spectrum of the Laplacian reflects several interesting properties of the associated graph, especially in the case of undirected graphs (see, for example, [10]).

$G(V, E, A)$ is *strongly connected* if it contains a directed path from any vertex j to any vertex l (i.e., a sequence of vertices starting with j and ending with l such that $(v_k, v_{k+1}) \in E$ for any two consecutive vertices v_k, v_{k+1}); G is *weakly connected* if there is such a path in the *associated undirected graph*, with adjacency matrix $A + A^T$.

For time-varying interconnections, a time-varying graph $G(t)$ is used and all the previously defined elements simply depend on time. If the elements of $A(t)$ are bounded and satisfy some threshold $a_{jk}(t) \geq \delta > 0$ $\forall (j,k) \in E(t)$ and $\forall t$, then $G(t)$ is called a *δ-digraph.* The present paper always considers δ-digraphs.

In a δ-digraph $G(V, E, A)$, vertex j is said to be *connected* to vertex k *across* $[t_1, t_2]$ if there is a path from j to k in the digraph $\bar{G}(V, \bar{E}, \bar{A})$ defined by

$$\bar{a}_{jk} = \begin{cases} \int_{t_1}^{t_2} a_{jk}(t)dt & \text{if } \int_{t_1}^{t_2} a_{jk}(t)dt \geq \delta, \\ 0 & \text{if } \int_{t_1}^{t_2} a_{jk}(t)dt < \delta, \end{cases}$$
$$(j,k) \in \bar{E} \text{ iff } \bar{a}_{jk} \neq 0 \, .$$

$\bar{G}$ can be seen as a time-integrated graph, while the δ-criterion prevents vanishing edges. A δ-digraph $G(t)$ is called *uniformly connected* if there exist a vertex k and a time horizon $T > 0$ such that $\forall t$, k is connected to all other vertices across $[t, t+T]$.

2.2. Specific manifolds. The concepts presented in this paper are illustrated on two particular manifolds, $SO(n)$ and $Grass(p, n)$.

The special orthogonal Lie group $SO(n)$. This can be viewed as the set of positively oriented orthonormal bases of $\mathbb{R}^n$, or equivalently, as the set of rotation matrices in $\mathbb{R}^n$; it is the natural state space for the orientation of a rigid body in $\mathbb{R}^n$. In its canonical representation, used in the present paper, a point of $SO(n)$ is characterized by a real $n \times n$ orthogonal matrix Q with determinant equal to $+1$. $SO(n)$ is homogeneous (as any Lie group), compact, and connected. It has dimension $n(n-1)/2$.

The Grassmann manifold $Grass(p, n)$. Each point on $Grass(p, n)$ denotes a p-dimensional subspace $\mathcal{Y}$ of $\mathbb{R}^n$. The dimension of $Grass(p, n)$ is $p(n-p)$. Since $Grass(n-p, n)$ is isomorphic to $Grass(p, n)$, by identifying orthogonally complementary subspaces, in this paper we can assume without loss of generality that $p \leq \frac{n}{2}$. For the special case $p = 1$, the Grassmann manifold $Grass(1, n)$ is also known as the *projective space* in dimension n. $Grass(p, n)$ is connected, compact, and homogeneous as the quotient of the orthogonal Lie group $O(n)$ by $O(p) \times O(n-p)$. Indeed, $\mathcal{Y} \in Grass(p, n)$ can be represented, for instance, by a (not necessarily positively oriented) orthonormal basis $Q \in O(n)$ whose first p column-vectors span $\mathcal{Y}$; the same point $\mathcal{Y} \in Grass(p, n)$ is represented by any Q whose first p column-vectors span $\mathcal{Y}$ ($O(p)$-symmetry) and whose last $n-p$ column-vectors span the orthogonal complement of $\mathcal{Y}$ ($O(n-p)$-symmetry). Other quotient structures for $Grass(p, n)$ are discussed in [1].

A matrix manifold representation of $Grass(p, n)$ found in [1] assigns to $\mathcal{Y}$ any $n \times p$ matrix Y of p orthonormal column-vectors spanning $\mathcal{Y}$ (p-basis representation); all Y corresponding to rotations and reflections of the p column-vectors in $\mathcal{Y}$ represent the same $\mathcal{Y}$ ($O(p)$-symmetry), so this representation is not unique. The dimension of this representation is $np - p(p+1)/2$. In [29], a point of $Grass(p, n)$ is represented by $\Pi = YY^T$, the orthonormal projector on $\mathcal{Y}$ (projector representation); using the orthonormal projector on the space orthogonal to $\mathcal{Y}$, $\Pi_\perp = I_n - YY^T$ where I_n denotes the $n \times n$ identity matrix, is strictly equivalent. The main advantage of this representation is that there exists a bijection between $Grass(p, n)$ and the orthonormal projectors of rank p such that the projector representation makes $Grass(p, n)$ an embedded submanifold of the cone $\mathbb{S}_n^+$ of $n \times n$ symmetric positive semidefinite matrices. A disadvantage of this representation is its large dimension $n(n+1)/2$.

3. The induced arithmetic mean. A homogeneous manifold $\mathcal{M}$ is a manifold with a transitive group action by a Lie group $\mathcal{G}$: it is isomorphic to the quotient manifold $\mathcal{G}/\mathcal{H}$ of a group $\mathcal{G}$ by one of its subgroups $\mathcal{H}$. Informally, it can be seen as a manifold on which "all points are equivalent." The present paper considers connected compact homogeneous manifolds satisfying the following embedding property.

ASSUMPTION 1. *$\mathcal{M}$ is a connected compact homogeneous manifold smoothly embedded in $\mathbb{R}^m$ with the Euclidean norm $\|y\| = r_{\mathcal{M}}$ constant over $y \in \mathcal{M}$. The Lie group $\mathcal{G}$ acts as a subgroup of the orthogonal group on $\mathbb{R}^m$.*

It is a well-known fact of differential geometry that any smooth $\frac{m}{2}$-dimensional Riemannian manifold can be smoothly embedded in $\mathbb{R}^m$. The additional condition $\|y\| = r_{\mathcal{M}}$ is in agreement with the fact that all points on $\mathcal{M}$ should be equivalent. It is sometimes preferred to represent $y \in \mathcal{M}$ by a matrix $B \in \mathbb{R}^{n_1 \times n_2}$ instead of a vector. Componentwise identification $\mathbb{R}^{n_1 \times n_2} \cong \mathbb{R}^m$ is assumed whenever necessary; the corresponding norm is the *Frobenius norm* $\|B\| = \sqrt{\text{trace}(B^T B)}$.

Consider a set of N agents on a manifold $\mathcal{M}$ satisfying Assumption 1. The position of agent k is denoted by y_k and its weight by w_k.

DEFINITION 3.1. *The* induced arithmetic mean *$IAM \subseteq \mathcal{M}$ of N agents of weights $w_k > 0$ and positions $y_k \in \mathcal{M}$, $k = 1, \ldots, N$, is the set of points in $\mathcal{M}$ that globally minimize the weighted sum of squared Euclidean distances in $\mathbb{R}^m$ to each y_k as follows:*

$$IAM = \underset{c \in \mathcal{M}}{\text{argmin}} \sum_{k=1}^{N} w_k \, d^2_{\mathbb{R}^m}(y_k, c) = \underset{c \in \mathcal{M}}{\text{argmin}} \sum_{k=1}^{N} w_k \, (y_k - c)^T (y_k - c) \,. \tag{3.1}$$

The anti induced arithmetic mean *$AIAM \subseteq \mathcal{M}$ is the set of points in $\mathcal{M}$ that globally maximize the weighted sum of squared Euclidean distances* in $\mathbb{R}^m$ *to each y_k as follows:*

$$AIAM = \underset{c \in \mathcal{M}}{\text{argmax}} \sum_{k=1}^{N} w_k \, d^2_{\mathbb{R}^m}(y_k, c) = \underset{c \in \mathcal{M}}{\text{argmax}} \sum_{k=1}^{N} w_k \, (y_k - c)^T (y_k - c) \,. \tag{3.2}$$

The terminology is derived from [31], where the *IAM* on $SO(3)$ is called the *projected arithmetic mean.* The point in Definition 3.1 is that distances are measured *in the embedding space* $\mathbb{R}^m$. It thereby differs from the canonical definition of mean of N agents on $\mathcal{M}$, the *Karcher mean* [24, 36, 16, 20], which uses the geodesic distance $d_{\mathcal{M}}$ along the Riemannian manifold $\mathcal{M}$ (with, in the present setting, the Riemannian metric induced by the embedding of $\mathcal{M}$ in $\mathbb{R}^m$) as follows:

$$C_{Karcher} = \underset{c \in \mathcal{M}}{\text{argmin}} \sum_{k=1}^{N} w_k \, d^2_{\mathcal{M}}(y_k, c) \,.$$

The IAM has the following properties:

1. The *IAM* of a single point y_1 is the point itself.

2. The *IAM* is invariant under permutations of agents of equal weights.
3. The *IAM* commutes with the symmetry group of the homogeneous manifold.
4. The *IAM* does not always reduce to a single point.

The last feature seems unavoidable for any mean (including the Karcher mean) that satisfies the other properties. The main advantage of the *IAM* over the Karcher mean is computational. The *IAM* and *AIAM* are closely related to the centroid.

Definition 3.2. *The* centroid $C_e \in \mathbb{R}^m$ *of* N *weighted agents located on* $\mathcal{M}$ *is*

$$C_e = \frac{1}{W}\sum_{k=1}^{N} w_k\, y_k \ , \quad \textit{where} \ \ W = \sum_{k=1}^{N} w_k \ .$$

Since $\|c\| = r_{\mathcal{M}}$ for $c \in \mathcal{M}$ by Assumption 1, equivalent definitions for the *IAM* and *AIAM* are

$$\text{(3.3)} \qquad \mathit{IAM} = \operatorname*{argmax}_{c\in\mathcal{M}}(c^T\, C_e) \qquad \text{and} \qquad \mathit{AIAM} = \operatorname*{argmax}_{c\in\mathcal{M}}(-c^T\, C_e) \ .$$

Hence, computing the *IAM* and *AIAM* just involves a search for the global maximizers of a linear function on $\mathbb{R}^m$ in a very regular search space $\mathcal{M}$. Local maximization methods would even suffice if the linear function had no maxima on $\mathcal{M}$ other than the global maxima. This is the case for any linear function on $SO(n)$ and $Grass(p,n)$ (see section 3.1) as well as the n-dimensional sphere S^n in $\mathbb{R}^{n+1}$. Not knowing whether this property holds for all manifolds satisfying Assumption 1, we formulate the following blanket assumption.

Assumption 2. *The local maxima of any linear function* $f(c) = c^T\, b$ *over* $c \in \mathcal{M}$*, with* b *fixed in* $\mathbb{R}^m$*, are all global maxima.*

3.1. Examples.

These examples exclusively consider the *IAM*; from (3.3), the conclusions for the *AIAM* are simply obtained by replacing C_e with $-C_e$.

The circle. The circle embedded in $\mathbb{R}^2$ with its center at the origin satisfies Assumptions 1 and 2. The *IAM* is simply the central projection of C_e onto the circle. Hence it corresponds to the whole circle if $C_e = 0$ and reduces to a single point in other situations. The *IAM* uses the chordal distance between points, while the Karcher mean would use arclength distance.

The special orthogonal group. The embedding of $SO(n)$ as orthogonal matrices $Q \in \mathbb{R}^{n\times n}$, $\det(Q) > 0$, satisfies Assumption 1 since $\|Q\| = \sqrt{\operatorname{trace}(Q^TQ)} = \sqrt{n}$. It also satisfies Assumption 2 (proof in section 6). $C_e = \sum_k Q_k$ is a general $n \times n$ matrix. The *IAM* is linked to the *polar decomposition* of C_e. Any matrix B can be decomposed into UR with U orthogonal and R symmetric positive semidefinite; R is always unique, and U is unique if B is nonsingular [7]. Each U is a global minimizer of $d_{\mathbb{R}^{n\times n}}(c, B)$ over $c \in O(n)$. Thus, if $\det(C_e) \geq 0$, the *IAM* contains all matrices U: $\det(U) > 0$ obtained from the polar decomposition of C_e; this was already noticed in [31]. When $\det(C_e) < 0$, the result is more complicated but still has a closed-form solution.

Proposition 3.3. *Consider* U *an orthogonal matrix obtained from the polar decomposition* $C_e = UR$*. The IAM of* N *points on* $SO(n)$ *is characterized as follows:*

1. *If* $\det(C_e) \geq 0$*, then IAM* $= \{U : \det(U) > 0\}$*. It reduces to a single point if the multiplicity of* 0 *as an eigenvalue of* C_e *is less or equal to* 1*.*

2. *If* $\det(C_e) \leq 0$*, then IAM* $= \{UHJH^T\}$*, where* $\det(U) < 0$*,* H *contains the orthonormalized eigenvectors of* R *with an eigenvector corresponding to the smallest eigenvalue of* R *in the first column, and* $\left(\begin{smallmatrix} -1 & 0 \\ 0 & I_{n-1} \end{smallmatrix}\right)$ *. The IAM reduces to a single point if the smallest eigenvalue of* R *has multiplicity* 1*.*

Proof. The proof is provided in section 6 after we introduce further necessary material to compute the critical points of $c^T C_e$, among which the local maxima are selected. □

The Grassmann manifold. The representation of $Grass(p,n)$ with p-bases Y_k is not an embedding and cannot be used in the proposed framework because the p-dimensional subspace of $\mathbb{R}^n$ spanned by the columns of $C_e = \sum_k Y_k$ would depend on the particular matrices Y_k chosen to represent the subspaces $\mathcal{Y}_k$. The *IAM* is defined with the projector representation, embedding $Grass(p,n)$ in $\mathbb{S}_n^+$. The latter satisfies Assumption 1; the Frobenius norm of a p-rank projector is $\sqrt{p}$. It also satisfies Assumption 2 (proof in section 6). The centroid C_e of N projectors is generally a symmetric positive semidefinite matrix of rank $\geq p$.

PROPOSITION 3.4. *The IAM contains all dominant p-eigenspaces of C_e. It reduces to a single point if the p largest and $(p{+}1)$-largest eigenvalues of C_e are different.*

Proof. The proof follows along the same lines as that for $SO(n)$, and is postponed to section 6. □

In fact, for $\mathcal{Y} \in Grass(p,n)$ with a p-basis Y and the projector $\Pi_{\mathcal{Y}} = YY^T$, the cost function in (3.3) becomes

$$f(\Pi_{\mathcal{Y}}) = \mathrm{trace}(\Pi_{\mathcal{Y}} C_e) = \mathrm{trace}(Y^T C_e Y) = \mathrm{trace}((Y^T Y)^{-1} Y^T C_e Y), \tag{3.4}$$

where the last expression is equal to the generalized Rayleigh quotient for the computation of the dominant p-eigenspace of C_e. The computation of eigenspaces from cost function (3.4) is extensively covered in [1, 2]. Furthermore, it is a well-known fact of linear algebra that the p largest eigenvalues (the others being 0) of $\Pi_{\mathcal{Y}}\Pi_k$ are the squared cosines of the principal angles ϕ_k^i, $i = 1, \dots, p$, between subspaces $\mathcal{Y}$ and $\mathcal{Y}_k$. This provides a geometrical meaning for the *IAM* of subspaces: it minimizes the sum of squared sines of principal angles between the set of subspaces $\mathcal{Y}_k$, $k = 1, \dots, N$, and a centroid candidate subspace $\mathcal{Y}$, i.e., *IAM* $= \mathrm{argmin}_{\mathcal{Y}} \sum_{k=1}^N \sum_{i=1}^p \sin^2(\phi_k^i)$. The Karcher mean admits the same formula with $\sin^2(\phi_k^i)$ replaced by $(\phi_k^i)^2$ [11].

4. Consensus.

Consider a set of agents with positions y_k, $k = 1, \dots, N$, on a manifold $\mathcal{M}$ satisfying Assumption 1. The rest of this paper assumes equal weights $w_k = 1\ \forall k$; extension to weighted agents is straightforward. Suppose that the agents are interconnected according to a fixed digraph G of adjacency matrix $A = [a_{jk}]$.

DEFINITION 4.1. Synchronization *is the configuration where $y_j = y_k\ \forall j,k$.*

DEFINITION 4.2. *A* consensus *configuration with graph G is a configuration where each agent k is located at a point of the IAM of its neighbors $j \rightsquigarrow k$, weighted according to the strength of the corresponding edge. Similarly, an* anticonsensus *configuration satisfies this definition with IAM replaced by AIAM.*

$$\text{Consensus:} \qquad y_k \in \underset{c\in\mathcal{M}}{\mathrm{argmax}} \left(c^T \sum_{j=1}^N a_{jk}\, y_j \right) \qquad \forall k\ . \tag{4.1}$$

$$\text{Anticonsensus:} \qquad y_k \in \underset{c\in\mathcal{M}}{\mathrm{argmin}} \left(c^T \sum_{j=1}^N a_{jk}\, y_j \right) \qquad \forall k\ . \tag{4.2}$$

Note that consensus is defined as a Nash equilibrium: each agent minimizes its cost function, *assuming the others fixed*; the possibility of decreasing cost functions by moving several agents simultaneously is not considered. Consensus is graph-dependent: agent k reaches consensus when it minimizes its distance to agents $j \rightsquigarrow k$.

PROPOSITION 4.3. *If G is an equally weighted complete graph, then the only possible consensus configuration is synchronization.*

Proof. At consensus the y_k satisfy $y_k^T \sum_{j \neq k} y_j \geq c^T \sum_{j \neq k} y_j \forall k$ and $\forall c \in \mathcal{M}$. Furthermore, it is obvious that $y_k^T y_k > y_k^T c$ for any $c \in \mathcal{M} \setminus \{y_k\}$. As a consequence, $y_k^T \sum_{j=1}^N y_j > c^T \sum_{j=1}^N y_j \;\; \forall c \in \mathcal{M} \setminus \{y_k\}$ and $\forall k$. Thus according to (3.3), each y_k is located at the *IAM* of all the agents, which moreover reduces to a single point; thus $y_k = y_j = \mathit{IAM}(\{y_l : l = 1, \ldots, N\}) \; \forall k, j$. □

Synchronization is a configuration of complete consensus. To similarly characterize a configuration of complete anticonsensus, it appears meaningful to require that the *IAM* of the agents be the entire manifold $\mathcal{M}$; this is called a *balanced* configuration.

DEFINITION 4.4. *N agents are* balanced *if their IAM contains all $\mathcal{M}$.*

Balancing implies some spreading of the agents on the manifold. A full characterization of balanced configurations seems complicated. Balanced configurations do not always exist (typically, when the number of agents is too small) and are mostly not unique (they can appear in qualitatively different forms). The following link exists between anticonsensus for the equally weighted complete graph and balancing.

PROPOSITION 4.5. *All balanced configurations are anticonsensus configurations for the equally weighted complete graph.*

Proof. For the equally weighted complete graph, (4.2) can be written

$$y_k \in \underset{c \in \mathcal{M}}{\operatorname{argmin}} \left(c^T \, (N \, C_e - y_k) \right) \qquad \forall k \, . \tag{4.3}$$

Assume that the agents are balanced. This means that $f(c) = c^T \, C_e$ must be constant over $c \in \mathcal{M}$. Therefore (4.3) reduces to $y_k = y_k \; \forall k$ which is trivially satisfied. □

In contrast to Proposition 4.3, Proposition 4.5 does not establish a necessary and sufficient condition; indeed, anticonsensus configurations for the equally weighted complete graph that are not balanced do exist, though they seem exceptional.

4.1. Examples. The following examples illustrate, among others, the last assertions about balanced configurations.

The circle. Anticonsensus configurations for the equally weighted complete graph are fully characterized in [42]. It is shown that the only anticonsensus configurations that are not balanced correspond to $(N + 1)/2$ agents at one position, and $(N - 1)/2$ agents at the opposite position on the circle, for N odd. Balanced configurations are unique for $N = 2$ and $N = 3$ and form a continuum for $N > 3$.

Another interesting illustration is the equally weighted *undirected ring graph* in which each agent is connected to two neighbors such that the graph forms a single closed undirected path. Regular consensus configurations correspond to situations with consecutive agents in the path always separated by the same angle $0 \leq \chi \leq \pi/2$; regular anticonsensus configurations have $\pi/2 \leq \chi \leq \pi$. In addition, for $N \geq 4$, irregular consensus and anticonsensus configurations exist where nonconsecutive angles of the regular configurations are replaced by $(\pi - \chi)$. As a consequence we have the following:

1. Several qualitatively different (anti)consensus configurations exist.
2. Consensus and anticonsensus configurations can be equivalent when discarding the graph. For example, the positions occupied by 7 agents separated by $2\pi/7$ (consensus) or $4\pi/7$ (anticonsensus) are strictly equivalent; the only difference, based on *which agent* is located at *which position*, concerns the way the links are drawn.

3. Degenerate configurations of simultaneous consensus and anticonsensus exist (e.g., $\chi = \pi/2$ for $N = 4, 8, \ldots$); this singularity is specific to the particular graph.

4. There is no common anticonsensus state for all ring graphs. Indeed, considering an agent k, a common anticonsensus state would require that any two other agents, as potential neighbors of k, are either separated by π or located at both sides of k at a distance of $\chi \geq \pi/2$; one easily verifies that this cannot be satisfied $\forall\ k$.

The special orthogonal group. Simulations of the algorithms proposed in this paper suggest that balanced configurations always exist for $N \geq 2$ if n is even and for $N \geq 4$ if n is odd. Under these conditions, convergence to an anticonsensus state that is not balanced is not observed for the equally weighted complete graph.

The Grassmann manifold. Balanced states on $Grass(p, n)$ appear if all eigenvalues of C_e are equal. Since $\text{trace}(C_e) = \frac{1}{N}\sum_k \text{trace}(\Pi_k) = p$, this requires $C_e = \frac{p}{n} I_n$. This is not always possible with N orthonormal projectors of rank p. As for $SO(n)$, simulations tend to indicate that it is possible when N is large enough; however, computing the minimal value of N for a given n and p is not straightforward.

5. Consensus optimization strategy. The presence of a maximization condition in the definitions of the previous sections naturally points to the use of optimization methods. The present section introduces a cost function whose optimization leads to (anti)consensus configurations. For a graph G with adjacency matrix $A = [a_{jk}]$ and associated Laplacian $L^{(i)} = [l^{(i)}_{jk}]$ and the variable $y = (y_1, \ldots, y_N) \in \mathcal{M}^N$, define

$$(5.1) \qquad P_L(y) \;=\; \frac{1}{2N^2}\sum_{k=1}^{N}\sum_{j=1}^{N} a_{jk}\; y_j^T y_k \;=\; \xi_1 - \frac{1}{4N^2}\sum_{k=1}^{N}\sum_{j=1}^{N} a_{jk}\; \|y_j - y_k\|^2$$

with constant $\xi_1 = \frac{r^2_{\mathcal{M}}}{4N^2}\sum_k \sum_{j=1}^N a_{jk}$. The index L refers to the fact that (5.1) can also be written as a quadratic form on the graph Laplacian as follows:

$$(5.2) \qquad P_L(y) \;=\; \xi_2 - \frac{1}{2N^2}\sum_{k=1}^{N}\sum_{j=1}^{N} l^{(i)}_{jk}\; y_j^T y_k \qquad \text{with constant } \xi_2 = \frac{r^2_{\mathcal{M}}}{2N^2}\sum_{k=1}^{N} d^{(i)}_k\ .$$

In [37] and [43], this form of P_L is studied on the circle for undirected equally weighted graphs. For the unit-weighted complete graph, $P := P_L + \frac{r^2_{\mathcal{M}}}{2N}$ equals

$$(5.3) \qquad P(y) = \frac{1}{2}\|C_e\|^2\ ,$$

proportional to the squared norm of the centroid C_e. This is a classical measure of the synchrony of phase variables on the circle S^1, used for decades in the literature on coupled oscillators; in the context of the Kuramoto model, $P(y)$ is known as the "complex order parameter" (because $\mathbb{R}^2$ is usually identified with $\mathbb{C}$ in that context). In [42], P is used to derive gradient algorithms for synchronization (by maximizing (5.3)) or balancing (by minimizing (5.3)) on S^1.

Proposition 5.1. *Synchronization of the N agents on $\mathcal{M}$ is the unique global maximum of P_L whenever the graph G associated with $L^{(i)}$ is weakly connected.*

Proof. According to the second form of (5.1), P_L reaches its global maximum when $y_j = y_k$ for all j, k for which $a_{jk} \neq 0$. If G is weakly connected, this equality propagates through the whole graph such that $y_1 = y_2 = \cdots = y_N$. □

Proposition 5.2. *Consider N agents on a manifold $\mathcal{M}$ satisfying Assumptions 1 and 2. Given an undirected graph G, a local maximum of the associated cost function*

$P_L(y)$ necessarily corresponds to a consensus configuration, and a local minimum of $P_L(y)$ necessarily corresponds to an anticonsensus configuration.

Proof. The proof is given for local maxima; it is strictly equivalent for local minima. For $y^* = (y_1^* \dots y_N^*)$ to be a local maximizer of P_L, y_k^* must be, for each k, a local maximizer of $p_k(c) := P_L(y_1^* \dots y_{k-1}^*,\, c,\, y_{k+1}^* \dots y_N^*)$. Since $A = A^T$, p_k takes the linear form $p_k(c) = \xi_k + \frac{1}{N^2} c^T \left(\sum_{j=1}^N a_{jk}\, y_j^*\right)$ with ξ_k constant $\forall k$. Thanks to Assumption 2, all local maxima of $p_k(c)$ are global maxima. Therefore, y_k^* is a global maximum of $p_k(c)$ $\forall$ k, which corresponds to Definition 4.2 of consensus. □

Proposition 5.2 establishes that a *sufficient* condition for (anti)consensus configurations is to optimize P_L. However, nothing guarantees that this is also *necessary.* In general, optimizing P_L will thus provide proven (anti)consensus configurations, but not necessarily all of them (this is because consensus maximizes P_L on $\mathcal{M}^N$ *for only moving one agent with others fixed*, and not along directions of *combined motion of several agents*). The remaining sections of this paper present algorithms that drive the swarm to (anti)consensus. Being based on the optimization of P_L, these algorithms do not necessarily target all possible (anti)consensus configurations. For instance, for a tree, maximization of P_L always leads to synchronization, although other consensus configurations can exist.

5.1. Examples. On $SO(n)$ and $Grass(p, n)$, P_L with matrix forms for the elements y_k becomes

$$P_L(y) \;=\; \frac{1}{2N^2}\sum_{j=1}^{N}\sum_{k=1}^{N} a_{jk}\,\mathrm{trace}(y_j^T y_k) \qquad \text{with } y_k \in \mathbb{R}^{n\times n}\ \forall k\ . \tag{5.4}$$

The special orthogonal group. Each term $Q_j^T Q_k = Q_j^{-1} Q_k$ is itself an element of $SO(n)$. It is actually the unique element of $SO(n)$ translating Q_j into Q_k by matrix (group) multiplication on the right. Hence, on the Lie group $SO(n)$, the order parameter P_L measures the sum of the traces of the elements translating connected agents into each other. Observing that the trace is maximal for the identity matrix and considering the particular case of $SO(2)$, one can easily imagine how the trace of $Q_j^{-1} Q_k$ characterizes the distance between Q_j and Q_k. This cost function has been previously used in [8, 34] as a measure of disagreement on $SO(3)$.

The Grassmann manifold. On $Grass(p, n)$, (5.4) can be rewritten as

$$P_L(\mathcal{Y}) = \frac{1}{2N^2}\sum_{j=1}^{N}\sum_{k=1}^{N} a_{jk}\left(\sum_{i=1}^{p}\cos^2(\phi_{jk}^i)\right)$$

with ϕ_{jk}^i = the ith principal angle between $\mathcal{Y}_j$ and $\mathcal{Y}_k$. This reformulation has previously appeared in [11, 4, 1].

6. Gradient consensus algorithms.

The previous sections pave the way for ascent and descent algorithms on P and P_L. This paper considers continuous-time gradient algorithms, but any descent or ascent algorithm—in particular, discrete-time—will achieve the same task; see [2] for extensive information on this subject. In the present paper, the gradient is always defined with the canonical metric induced by the embedding of $\mathcal{M}$ in $\mathbb{R}^m$.

6.1. Fixed undirected graphs.

A gradient algorithm for P_L leads to

$$\dot{y}_k(t) = 2N^2\alpha\ \mathrm{grad}_{k,\mathcal{M}}(P_L)\ , \qquad k = 1, \dots, N\ , \tag{6.1}$$

where $\alpha > 0$ (resp., $\alpha < 0$) for consensus (resp., anticonsensus), $\dot{y}_k$ denotes the time-derivative of agent k's position, and $\mathrm{grad}_{k,\mathcal{M}}(f)$ denotes the gradient of f with respect to y_k along $\mathcal{M}$. This gradient can be obtained from the gradient in $\mathbb{R}^m$,

$$\mathrm{grad}_{k,\mathbb{R}^m}(P_L) = \frac{1}{2N^2}\sum_{j=1}^{N}(a_{jk} + a_{kj})\, y_j \ ,$$

by orthogonal projection $\mathrm{Proj}_{T\mathcal{M},k}$ onto the tangent space to $\mathcal{M}$ at y_k, yielding $\forall k$,
(6.2)

$$\dot{y}_k(t) = \alpha\ \mathrm{Proj}_{T\mathcal{M},k}\left(\sum_{j=1}^{N}(a_{jk} + a_{kj})y_j\right) = \alpha\ \mathrm{Proj}_{T\mathcal{M},k}\left(\sum_{j=1}^{N}(a_{jk} + a_{kj})(y_j - y_k)\right) .$$

The last equality comes from $\mathrm{Proj}_{T\mathcal{M},k}(y_k) = 0$. It shows that to implement this consensus algorithm, each agent k must know the relative position with respect to itself of all agents j such that $j \rightsquigarrow k$ or $k \rightsquigarrow j$. Since the information flow is restricted to $j \rightsquigarrow k$, (6.2) can be implemented only for undirected graphs, for which it becomes

$$\dot{y}_k(t) = 2\alpha\ \mathrm{Proj}_{T\mathcal{M},k}\left(\sum_{j=1}^{N} a_{jk}(y_j - y_k)\right) , \qquad k = 1, \dots, N \ . \tag{6.3}$$

In the special case of a complete unit-weighted graph,

$$\dot{y}_k(t) = 2\alpha N\ \mathrm{Proj}_{T\mathcal{M},k}\left(C_e(t) - y_k\right) \ , \qquad k = 1, \dots, N \ . \tag{6.4}$$

PROPOSITION 6.1. *A group of N agents moving according to* (6.3) *on a manifold $\mathcal{M}$ satisfying Assumptions* 1 *and* 2*, where the graph G associated to $A = [a_{jk}]$ is undirected, always converges to a set of equilibrium points. If $\alpha < 0$, all asymptotically stable equilibria are anticonsensus configurations for G. If $\alpha > 0$, all asymptotically stable equilibria are consensus configurations for G (in particular, for the equally weighted complete graph, the only asymptotically stable configuration is synchronization).*

Proof. $\mathcal{M}$ being compact and the a_{jk} bounded, P_L is upper- and lower-bounded. P_L is always increasing (decreasing) for $\alpha > 0$ ($\alpha < 0$) along solutions of (6.3), since

$$\dot{P}_L \ = \ \sum_{k=1}^{N} \dot{y}_k^T\, \mathrm{grad}_{k,\mathcal{M}}(P_L) \ = \ 2N^2\alpha \sum_{k=1}^{N} \|\mathrm{grad}_{k,\mathcal{M}}(P_L)\|^2 \ .$$

By LaSalle's invariance principle, the swarm converges towards a set where $\dot{P}_L = 0$, implying $\mathrm{grad}_{k,\mathcal{M}}(P_L) = 0 \Leftrightarrow \dot{y}_k = 0\ \forall k$, and the swarm converges to a set of equilibria. For $\alpha > 0$ ($\alpha < 0$), since P_L always increases (decreases) along solutions, only local maxima (minima) can be asymptotically stable. Proposition 5.2 states that all local maxima (minima) of P_L correspond to consensus (anticonsensus). □

Remark 1. Computing $\mathrm{grad}_{k,\mathcal{M}}$ directly along the manifold, as in [2], can be much more efficient if the dimension of $\mathcal{M}$ is substantially lower than m (see section 6.3).

6.2. Extension to directed and time-varying graphs. Formally, algorithm (6.3) can be written for directed and even time-varying graphs, although the gradient property is lost for directed graphs and has no meaning in the time-varying case (since P_L then explicitly depends on time). Nevertheless, the general case of (6.3) with varying and directed graphs still exhibits synchronization properties.

It can be shown that synchronization is still a stable equilibrium; it is asymptotically stable if disconnected graph sequences are excluded. Its basin of attraction includes the configurations where all the agents are located in a convex set of $\mathcal{M}$. Indeed, convergence results on Euclidean spaces can be adapted to manifolds when agents are located in a convex set (see, e.g., [33]). On the other hand, examples where algorithm (6.3) with $\alpha > 0$ runs into a limit cycle can be built for cases as simple as undirected equally weighted (but varying) graphs on the circle (see section 6.3).

Simulations on $SO(n)$ and $Grass(p, n)$ seem to indicate that for randomly generated digraph sequences,[1] the swarm eventually converges to synchronization when $\alpha > 0$; this would correspond to *generic* convergence for unconstrained graphs.

Algorithm (6.3) can lead to a generalization of Vicsek's phase update law (see [48]) to manifolds. The Vicsek model is a discrete-time algorithm governing the headings of particles in the plane, and hence operates on the circle. It can be written as

$$(6.5) \qquad y_k(t+1) \in IAM\left(\{y_j(t) | j \rightsquigarrow k \text{ in } G(t)\} \cup \{y_k(t)\}\right), \qquad k = 1, \ldots, N,$$

with the definitions introduced in the present paper; interconnections among particles depend on their relative positions in the plane (so-called "proximity graphs"). Vicsek's law can be directly generalized in the form (6.5) to any manifold satisfying Assumption 1. Based on the previous discussions, it is clear why (6.5) can be viewed as a discrete-time variant of (6.3). When run asynchronously on a fixed undirected graph, (6.5) is an ascent algorithm for P_L; see [37] for a precise relationship between the continuous-time and discrete-time consensus algorithms on the circle.

6.3. Examples. Consensus on the circle is studied in [42, 37, 39, 43]; the other algorithms presented here are original.

The circle. Denoting angular positions by θ_k, the specific form of (6.3) for S^1 is

$$(6.6) \qquad \dot{\theta}_k = \alpha' \sum_{j=1}^{N} a_{jk} \sin(\theta_k - \theta_j), \qquad k = 1, \ldots, N.$$

For the equally weighted complete graph, this is strictly equivalent to the Kuramoto model [26] with identical (zero) natural frequencies.

Algorithm (6.6) can run into a limit cycle for varying graphs. Consider a regular consensus state for an equally weighted ring graph G_1, with consecutive agents separated by $\chi < \pi/2$ (local maximum of P_{L_1}). Define G_2 by connecting each agent to the agents located at an angle $\psi > \pi/2$ from itself with ψ properly fixed. G_2 is a collection of disconnected ring graphs and the swarm is at a local minimum of P_{L_2}. Starting the system in the neighborhood of that state and regularly switching between G_1 and G_2, the system will oscillate in its neighborhood, being driven away by G_2 and brought back by G_1 if consensus is intended, and conreversely if anticonsensus is intended.

[1] More precisely, the following distribution was examined: initially, each element a_{jk} independently takes a value in $\{0, 1\}$ according to a probability Prob$(1) = p$. The corresponding graph remains for a time t_{graph} uniformly distributed in $[t_{min}, t_{max}]$, after which a new graph is built as initially.

The special orthogonal group. The tangent space to $SO(n)$ at the identity I_n is the space of skew-symetric $n \times n$ matrices. By group multiplication, the projection of $B \in \mathbb{R}^{n\times n}$ onto the tangent space to $SO(n)$ at Q_k is $Q_k\,\mathrm{Skew}(Q_k^{-1}B) = Q_k\,(\frac{Q_k^T B}{2} - \frac{B^T Q_k}{2})$. This leads to the following explicit form of algorithm (6.3) on $SO(n)$, where the right-hand side depends only on relative positions of the agents with respect to k:

$$Q_k^{-1}\dot{Q}_k = \alpha \sum_{j=1}^{N} a_{jk}\ \left(Q_k^{-1}Q_j - Q_j^{-1}Q_k\right)\ , \qquad k = 1,\ldots,N\ . \tag{6.7}$$

Using Lemma A.1 in the appendix, the following proves that $SO(n)$ satisfies Assumption 2. It also includes the proof of Proposition 3.3.

PROPOSITION 6.2. *The manifold $SO(n)$ satisfies Assumption* 2.

Proof of Propositions 6.2 *and* 3.3. Consider a linear function $f(Q) = \mathrm{trace}(Q^T B)$ with $Q \in SO(n)$ and $B \in \mathbb{R}^{n\times n}$; $\mathrm{grad}_{\mathbb{R}^{n\times n}}(f) = B$, so $\mathrm{grad}_{SO(n)}(f) = \frac{Q}{2}(Q^T B - B^T Q)$. Since Q is invertible, critical points of f satisfy $(Q^T B - B^T Q) = 0$, meaning that they take the form described by Lemma A.1. Using notations of Lemma A.1, write $R = H\Lambda H^T$, where Λ contains the (nonnegative) eigenvalues of R. This leads to

$$Q = UHJH^T \quad \Rightarrow \quad Q^T B = HJ\Lambda H^T \quad \Rightarrow \quad f(Q) = -\sum_{j=1}^{l} \Lambda_{jj} + \sum_{j=l+1}^{n} \Lambda_{jj}\ .$$

If $l \geq 2$, select any $m \in [2, l]$ and define $Q_\varepsilon = UHJAH^T$, where A is the identity matrix, except that $A(1,1) = A(m,m) = \cos(\varepsilon)$ and $A(1,m) = -A(m,1) = \sin(\varepsilon)$ with ε arbitrarily small. It is straightforward to see that $f(Q_\varepsilon) > f(Q)$, unless $\Lambda_{11} = \Lambda_{mm} = 0$. Similarly, if $l = 1$ and $\exists\ m \geq 2$ such that $\Lambda_{mm} < \Lambda_{11}$, then $f(Q_\varepsilon) > f(Q)$ with Q_ε and A as defined previously. Therefore,

1. if $\det(B) \geq 0$, local maxima require $l = 0$ such that $Q = U$ and $f(Q)$ is the sum of the eigenvalues of R;

2. if $\det(B) \leq 0$, local maxima require U to take the form of Lemma A.1 with $l = 1$ and $\Lambda_{11} \leq \Lambda_{mm}\ \forall m$; thus the first column of H corresponds to a smallest eigenvalue of R, and $f(Q)$ is the sum of $n-1$ largest eigenvalues minus the smallest one.

This shows that all maxima of $f(Q)$ are global maxima (since they all take the same value) and, with $B = C_e$, characterizes the *IAM*. □

The Grassmann manifold. The projection of a matrix $M \in \mathbb{S}_n^+$ onto the tangent space to $Grass(p,n)$ at Π_k is given in [29] as $\Pi_k M \Pi_{\perp k} + \Pi_{\perp k} M \Pi_k$. This leads to

$$\dot{\Pi}_k = 2\alpha \sum_{j=1}^{N} a_{jk}\,(\Pi_k \Pi_j \Pi_{\perp k} + \Pi_{\perp k}\Pi_j \Pi_k)\ , \qquad k = 1,\ldots,N\ . \tag{6.8}$$

In practice, the basis representation Y_k is handier than Π_k since it involves smaller matrices. Computing the gradient of $P_L(\{\Pi_k\ ,\ k = 1,\ldots,N\}) = P_L(\{Y_k\,Y_k^T\ ,\ k = 1,\ldots,N\})$ directly on the quotient manifold as explained in [1] leads to the algorithm

$$\dot{Y}_k = 4\alpha \sum_{j=1}^{N} a_{jk}\ \left(Y_j\,M_{j\cdot k} - Y_k\,M_{j\cdot k}^T M_{j\cdot k}\right)\ , \qquad k = 1,\ldots,N\ , \tag{6.9}$$

where the $p \times p$ matrices $M_{j\cdot k}$ are defined as $M_{j\cdot k} = Y_j^T Y_k$. For theoretical purposes, the projector representation is an easier choice, as for the following proofs.

Proposition 6.3. *The Grassmann manifold satisfies Assumption* 2.

Proof of Propositions 6.3 *and* 3.4. Consider a linear function $f(\Pi) = \text{trace}(\Pi^T B)$, where $B \in \mathbb{S}_n^+$ and Π represents $\mathcal{Y} \in Grass(p, n)$; $\text{grad}_{\mathbb{R}^{n\times n}}(f) = B$, so $\text{grad}_{Grass(p,n)}(f) = \Pi B \Pi_\perp + \Pi_\perp B \Pi$. The ranges of the first and second terms in $\text{grad}_{Grass(p,n)}(f)$ are at most $\mathcal{Y}$ and its orthogonal complement, respectively, so they both equal zero at a critical point $\mathcal{Y}^*$, such that $\mathcal{Y}^*$ is an invariant subspace of B. In an appropriate basis $(e_1 \dots e_n)$, write $\Pi^* = \text{diag}(1, \dots, 1, 0, \dots, 0)$ and $B = \text{diag}(\mu_1, \dots, \mu_p, \mu_{p+1}, \dots, \mu_n)$. If $\exists\, d \leq p$ and $l > p$ such that $\mu_d < \mu_l$, then any variation of Π^* rotating e_d towards e_l strictly increases $f(\Pi)$. Therefore, at local maxima of $f(\Pi)$, the p-dimensional space corresponding to Π must be an eigenspace of B corresponding to p largest eigenvalues. This implies that at any local maximum, $f(\Pi)$ equals the sum of p largest eigenvalues of B, so Assumption 2 is satisfied. Replacing B by C_e proves Proposition 3.4. □

7. Consensus algorithms with estimator variables.

Section 6 derives algorithms that lead to a consensus situation linked to the interconnection graph. But in many applications, the interconnection graph is just a restriction on communication possibilities, under which one actually wants to achieve a consensus for the complete graph. Moreover, allowing directed and time-varying communication graphs is desirable for robustness. This section presents algorithms achieving the same performance as those of section 6 for the equally weighted complete graph—that is, driving the swarm to synchronization or to a subset of the anticonsensus configurations for the equally weighted complete graph which seems to contain little more than balancing—under very weak conditions on the actual communication graph. However, this reduction of information channels must be compensated by adding a consensus variable $x_k \in \mathbb{R}^m$, which interconnected agents are able to communicate, to the state space of each agent.

7.1. Synchronization algorithm.

For synchronization purposes, the agents run a consensus algorithm on their estimator variables x_k in $\mathbb{R}^m$, $k = 1, \dots, N$, initialized arbitrarily but independently and such that they can take any value in an open subset of $\mathbb{R}^m$; $\forall k$, agent k's position y_k on $\mathcal{M}$ independently tracks (the projection on $\mathcal{M}$ of) x_k. This leads to

$$(7.1)\quad \dot{x}_k = \beta \sum_{j=1}^{N} a_{jk}\,(x_j - x_k), \quad \beta > 0,$$

$$(7.2)\quad \dot{y}_k = \gamma_S\, \text{grad}_{k,\mathcal{M}}(y_k^T x_k) = \gamma_S\, \text{Proj}_{T\mathcal{M},k}(x_k), \quad \gamma_S > 0\ , \quad k = 1, \dots, N.$$

Equation (7.1) is a classical consensus algorithm in $\mathbb{R}^m$, where $\dot{x}_k(t)$ points from $x_k(t)$ towards the centroid of the (appropriately weighted) $x_j(t)$ for which $j \rightsquigarrow k$ at time t. According to [33, 32, 35], if the time-varying communication graph $G(t)$ is piecewise continuous in time and uniformly connected, then all the x_k exponentially converge to a common consensus value x_∞; moreover, if $G(t)$ is balanced for all t, then $x_\infty = \frac{1}{N}\sum_k x_k(0)$ (i.e., x_∞ is the centroid of the initial x_k). This implies the following convergence property for (7.1), (7.2), where the notation *IAM*$_g$ generalizes the definition (3.3) of the *IAM* when the points defining C_e are not on $\mathcal{M}$.

Proposition 7.1. *Consider a piecewise continuous and uniformly connected graph $G(t)$ and a manifold $\mathcal{M}$ satisfying Assumptions* 1 *and* 2*. The only stable limit configuration of the y_k under* (7.1), (7.2)*, with the x_k initialized arbitrarily but independently and such that they can take any value in an open subset of* $\mathbb{R}^m$*, is synchronization at* $y_\infty = \text{Proj}_{T\mathcal{M},k}(x_\infty)$*; if $G(t)$ is balanced,* $y_\infty = \textit{IAM}_g\{x_k(0),\, k = 1, \dots, N\}$.

Proof. Convergence of (7.1) towards $x_k = x_\infty$ $\forall k$ is proved in [32]; the property $x_\infty = \frac{1}{N}\sum_k x_k(0)$ for balanced graphs is easy to check (see [35]). As a consequence, the asymptotic form of (7.1), (7.2) is a set of N independent systems,

$$x_k = x_\infty, \tag{7.3}$$

$$\dot{y}_k = \gamma_S \operatorname{Proj}_{T\mathcal{M},k}(x_\infty) \ , \qquad k = 1, \dots, N \ , \tag{7.4}$$

where x_∞ is a constant. According to [30], the ω-limit sets of the original system (7.1), (7.2) correspond to the chain recurrent sets of the asymptotic system (7.3), (7.4). The first equation is trivial. According to Proposition 4 in [21] and Sard's theorem, since (7.4) is a gradient ascent algorithm for $f(y_k) = y_k^T x_\infty$ and $f(y_k)$ is smooth (as the restriction of a smooth function to the smooth embedded manifold $\mathcal{M}$), the chain recurrent set of (7.4) is equal to its critical points. Since x_∞ is a linear combination of the $x_k(0)$, variations of the $x_k(0)$ are equivalent to variations of x_∞.

PROPERTY o. *Any open neighborhood O of any point $x_o \in \mathbb{R}^m$ contains a point x_a for which $f(y_k)$ has a unique* (*local = global, by Assumption* 2) *maximizer.*

Because of Property o, with respect to variations of the x_k, the situation "$f(y_k)$ has multiple maximizers" is unstable. The situation "$f(y_k)$ has a unique maximizer" is stable since it corresponds to a nonempty open set in $\mathbb{R}^m$; thus a convex neighborhood of x_∞ can be found in which the $x_k(t)$ will stay, by convexity of (7.1), and where $f(y_k)$ has a unique maximizer. With respect to variations of the y_k, the (thus unique) maximizer of $y_k^T x_k$ is the only stable equilibrium for gradient ascent algorithm (7.2) such that for $x_k \to x_\infty$ the only stable situation is synchronization.

Proof of Property o. If $y_k^T x_o$ has multiple maximizers, select one of them and call it y_*. Then for $\sigma > 0$, $y_k^T x_o + \sigma\, y_k^T y_* \leq y_*^T x_o + \sigma\, y_k^T y_* \leq y_*^T x_o + \sigma\, y_*^T y_*$ with equality holding if and only if $y_k = y_*$, so y_* is the unique maximizer of $y_k^T(x_o + \sigma y_*)$. Since any open neighborhood O of x_o contains points of the form $x_a = x_o + \sigma y_*$, $\sigma > 0$, property o is proved. □

7.2. Anticonsensus algorithm. For anticonsensus, in analogy with the previous section, each y_k evolves according to a gradient algorithm to maximize its distance to $x_k(t)$. If $x_k(t)$ asymptotically converges to $C_e(t)$, this becomes equivalent to the gradient anticonsensus algorithm (6.4). Imposing $x_k(0) = y_k(0)$ $\forall k$, the following algorithm achieves this purpose when $G(t)$ is balanced $\forall t$:

$$\dot{x}_k \ = \beta \sum_{j=1}^{N} a_{jk}\,(x_j - x_k) \ + \dot{y}_k, \quad \beta > 0, \tag{7.5}$$

$$\dot{y}_k \ = \gamma_B \operatorname{grad}_{k,\mathcal{M}}(y_k^T x_k) \ = \ \gamma_B \operatorname{Proj}_{T\mathcal{M},k}(x_k), \quad \gamma_B < 0 \ , \quad k = 1, \dots, N \ . \tag{7.6}$$

Note that the variables x_k and y_k are fully coupled; in a discrete-time version of this system, this essential feature of the algorithm must be retained in the form of *implicit* update equations in order to ensure convergence (see [39] for details).

PROPOSITION 7.2. *Consider a piecewise continuous, uniformly connected, and balanced graph $G(t)$ and a manifold $\mathcal{M}$ satisfying Assumptions* 1 *and* 2*. Then, algorithm* (7.5), (7.6) *with initial conditions $x_k(0) = y_k(0)$ $\forall k$ converges to an equilibrium configuration of the anticonsensus algorithm for the equally weighted complete graph, that is,* (6.4) *with $\alpha < 0$.*

Proof. First, show that $\frac{1}{N}\sum_k x_k(t) = \frac{1}{N}\sum_k y_k(t) = C_e(t)$. Since $x_k(0) = y_k(0)$ $\forall k$, it is true for $t = 0$. Thus it remains to show that $\sum_k \dot{x}_k(t) = \sum_k \dot{y}_k(t)$. This is the case because a balanced graph ensures that the first two terms on the right-hand side of the following expression cancel each other:

$$\sum_{k=1}^{N} \dot{x}_k(t) \;=\; \beta \sum_{j=1}^{N} \left(\sum_{k=1}^{N} a_{jk} \right) x_j \;-\; \beta \sum_{k=1}^{N} \left(\sum_{j=1}^{N} a_{jk} \right) x_k \;+\; \sum_{k=1}^{N} \dot{y}_k(t) \;.$$

Next, prove that $\forall k$, $\dot{y}_k(t)$ is a uniformly continuous function in $L_2(0,+\infty)$ such that Barbalat's lemma implies $\dot{y}_k \to 0$. First, show that $W(t) = \frac{1}{2}\sum_k x_k(t)^T x_k(t)$ is never increasing along the solutions of (7.5), (7.6). Denoting by $(x)_j$, $j = 1, \ldots, m$, the vectors of length N containing the jth component of every x_k, $k = 1, \ldots, N$, and by $L^{(i)}$ the in-Laplacian of the varying graph associated to the a_{jk}, we obtain

$$\dot{W}(t) \;=\; \sum_{k=1}^{N} x_k^T \dot{x}_k \;=\; \sum_{k=1}^{N} x_k^T \dot{y}_k \;-\; \beta \sum_{j=1}^{N} (x)_j^T L^{(i)} (x)_j \;.$$

The term containing $L^{(i)}$ is nonpositive because the Laplacian of balanced graphs is positive semidefinite (see [49]). Replacing $\dot{y}_k$ from (7.6) and noting that $x_k^T \mathrm{Proj}_{T\mathcal{M},k}(x_k) = (\mathrm{Proj}_{T\mathcal{M},k}(x_k))^T \mathrm{Proj}_{T\mathcal{M},k}(x_k)$, one obtains

$$\dot{W}(t) \;=\; \gamma_B \sum_{k=1}^{N} \|\mathrm{Proj}_{T\mathcal{M},k}(x_k)\|^2 \;-\; \beta \sum_{j=1}^{N} (x)_j^T L^{(i)} (x)_j \leq 0 \;. \tag{7.7}$$

Thus $W(t) \leq W(0) = \frac{N}{2}\, r_{\mathcal{M}}^2$, which implies that each $\dot{y}_k(t)$ is in $L_2(0,+\infty)$ since

$$\frac{1}{|\gamma_B|} \sum_{k=1}^{N} \int_0^{+\infty} \|\dot{y}_k(t)\|^2 \, dt \;\leq\; -\int_0^{+\infty} \dot{W}(t)\, dt \;\leq\; \frac{N}{2}\, r_{\mathcal{M}}^2 \;.$$

$W(t) \leq W(0)$ also implies that x_k is uniformly bounded $\forall k$; from (7.6), $\dot{y}_k$ is uniformly bounded as well. Combining these two observations, with the a_{jk} bounded, (7.5) shows that x_k has a bounded derivative, and hence is Lipschitz in t $\forall k$. Now write

$$\begin{aligned} &\|\dot{y}_k(x_k(t_1), y_k(t_1)) - \dot{y}_k(x_k(t_2), y_k(t_2))\| \\ &\;\leq \|\dot{y}_k(x_k(t_1), y_k(t_1)) - \dot{y}_k(x_k(t_2), y_k(t_1))\| + \|\dot{y}_k(x_k(t_2), y_k(t_1)) - \dot{y}_k(x_k(t_2), y_k(t_2))\| \,. \end{aligned}$$

The first term on the second line is bounded by $r_1\,|t_1 - t_2|$ for some r_1 since $\dot{y}_k$ is linear in x_k, and x_k is Lipschitz in t. The second term on the second line is bounded by $r_2\,|t_1 - t_2|$ for some r_2 since $\dot{y}_k$ is Lipschitz in y_k (as the gradient of a smooth function along the smooth manifold $\mathcal{M}$), and $\frac{d}{dt}(y_k) = \dot{y}_k$ is uniformly bounded. Hence, $\dot{y}_k$ is Lipschitz in t, and therefore uniformly continuous in t, such that Barbalat's lemma can be applied. Therefore $\dot{y}_k \to 0$. Thus from [30], the ω-limit sets of (7.1), (7.2) correspond to the chain recurrent sets of the asymptotic system

$$\begin{aligned} \dot{x}_k \;&=\; \beta \sum_{j=1}^{N} a_{jk}\,(x_j - x_k), \\ 0 \;&=\; \gamma_B \,\mathrm{Proj}_{T\mathcal{M},k}(x_k) \;. \end{aligned}$$

The second line is just a static condition. The chain recurrent set of the linear consensus algorithm in the first line reduces to its equilibrium set $x_k = x_\infty$ $\forall k$. But then, from the beginning of the proof, $x_k = C_e \; \forall k$ such that the static condition becomes $0 = \gamma_B \operatorname{Proj}_{T\mathcal{M},k}(C_e) \; \forall k$. This is the condition for an equilibrium of anticonsensus algorithm (6.4) with $\gamma_B = 2\alpha N$. □

In simulations, a swarm applying (7.5), (7.6) with $x_k(0) = y_k(0) \; \forall k$ seems to generically converge to an anticonsensus configuration of the equally weighted complete graph, that is, a *stable* equilibrium configuration of (6.4) with $\alpha < 0$.

7.3. Examples. Applying this strategy to the circle yields the results of [39]; the x_k reduce to vectors of $\mathbb{R}^2$ and algorithms (7.2) and (7.6), respectively, drive the y_k towards and away from the central projection of x_k onto the unit circle.

The special orthogonal and Grassmann manifolds. The particular balancing algorithms will not be detailed, as they are directly obtained from their synchronization counterparts. Introducing auxiliary $n \times n$-matrices X_k, (7.1) may be transcribed verbatim. Using previously presented expressions for $\operatorname{Proj}_{T\mathcal{M},k}(X_k)$, (7.2) becomes

$$\text{(7.8)} \qquad \text{On } SO(n): \quad Q_k^{-1}\dot{Q}_k \;=\; \frac{\gamma_S}{2}\left(Q_k^T X_k - X_k^T Q_k\right), \quad k = 1, \ldots, N,$$

$$\text{(7.9)} \qquad \text{On } Grass(p,n): \quad \dot{\Pi}_k \;=\; \gamma_S\left(\Pi_k X_k \Pi_{\perp k} + \Pi_{\perp k} X_k \Pi_k\right), \quad k = 1, \ldots, N.$$

Note that for $Grass(p, n)$, the projector representation must be used in (7.1) and (7.5) such that using $n \times n$ matrices X_k becomes unavoidable.

7.4. Remark on the communication of estimator variables. To implement the algorithms of this section, interconnected agents must communicate the values of their estimator variable x_k. It is important to note that the variables x_k may not be just a set of abstract scalars for each agent k; since x_k interacts with the geometric y_k, it must be a geometric quantity too. However, the x_k evolve in $\mathbb{R}^m$, while the original system lives on $\mathcal{M}$; the relative position of agents on $\mathcal{M}$ is a meaningful measurement, but nothing ensures a priori that a similar thing can be done in $\mathbb{R}^m$. A solution could be to use a common (thus external) reference frame in $\mathbb{R}^m$ and transmit the coordinates of the x_k in this frame. That solution would unfortunately imply that the swarm loses its full autonomy; however, the external frame is just used for "translation" purposes and does not interfere with the dynamics of the system.

When $\mathcal{M}$ is (a subgroup of) $SO(n)$, the algorithms can be reformulated such that they work completely autonomously if interconnected agents measure their relative positions $Q_k^T Q_j$. Indeed, define $Z_k = Q_k^T X_k$. Then (7.1), (7.2), for instance, becomes

$$\text{(7.10)} \qquad \dot{Z}_k \;=\; (Q_k^T \dot{Q}_k)^T Z_k \;+\; \beta \sum_{j=1}^{N} a_{jk}\left((Q_k^T Q_j) Z_j - Z_k\right),$$

$$\text{(7.11)} \qquad Q_k^T \dot{Q}_k \;=\; \frac{\gamma_S}{2}\left(Z_k - Z_k^T\right), \; k = 1, \ldots, N.$$

In this formulation, each agent k can represent Z_k as an array of scalars, whose columns express the column-vectors of X_k as coordinates in a local frame attached to k (i.e., in a frame rotated by Q_k with respect to a hypothetical reference frame). Premultiplying Z_j by $Q_k^T Q_j$ expresses X_j in the local frame of k, and $Q_k^T \dot{Q}_k$ expresses the velocity of Q_k (with respect to a hypothetical fixed reference) in the local frame of k as well. Thus (7.10), (7.11) actually corresponds to (7.1), (7.2) written in the local frame of k. Each agent k gets from its neighbors $j \rightsquigarrow k$ their relative positions $Q_k^T Q_j$ and the $n \times n$ arrays of numbers Z_j; from this it computes the update $\dot{Z}_k$ to its own array of numbers Z_k and the move it has to make with respect to its current position, $Q_k^T \dot{Q}_k$. The same can be done for the anticonsensus algorithm.

8. Conclusion.

The present paper makes three main contributions.

First, it defines the induced arithmetic mean of N points on an embedded connected compact homogeneous manifold $\mathcal{M}$; though it differs from the traditional Karcher mean, it has a clear geometric meaning with the advantage of being easily computable—see analytical solutions for $SO(n)$ and $Grass(p, n)$.

Second, a definition of consensus directly linked to the induced arithmetic mean is presented for these manifolds. In particular, the notion of balancing introduced in [42] for the circle is extended to connected compact homogeneous manifolds. Consensus for the equally weighted complete graph is equivalent to synchronization. Likewise, it appears in simulations that anticonsensus for the equally weighted complete graph leads to balancing (if N is large enough) even though this could not be proved.

Third, consensus is formulated as an optimization problem, and distributed consensus algorithms are designed for N agents moving on a connected compact homogeneous manifold. In a first step, gradient algorithms are derived for fixed undirected interconnection graphs; (anti)consensus configurations are their only stable equilibria. Similar algorithms are considered when the graph is allowed to be directed and/or to vary, but their convergence properties are mostly open. In a second step, the algorithms are modified by incorporating an estimator variable for each agent. In this setting, convergence to the (anti)consensus states of the equally weighted complete graph can be established theoretically for time-varying and directed interconnection graphs. The meaningful way of communicating estimators between agents remains an open issue when $\mathcal{M}$ is not a subgroup of $SO(n)$.

Running examples $SO(n)$ and $Grass(p, n)$ illustrate the validity of the discussion and provide geometric insight. The models and results obtained by applying this framework to the circle are strictly equivalent to existing models and results (most significantly in [42, 43, 39]). This draws a link from the present discussion to the vast literature about synchronization and balancing on the circle.

Appendix.

LEMMA A.1. *If $g(Q) = Q^T B - B^T Q$ with $Q \in SO(n)$ and $B \in \mathbb{R}^{n \times n}$, then $g(Q) = 0$ iff $Q = UHJH^T$, where $B = UR$ is a polar decomposition of B, the columns of H contain (orthonormalized) eigenvectors of R, and*

$$J = \begin{pmatrix} -I_l & 0 \\ 0 & I_{n-l} \end{pmatrix}, \qquad \begin{array}{ll} l \textit{ even} & \textit{if } \det(U) > 0, \\ l \textit{ odd} & \textit{if } \det(U) < 0. \end{array}$$

Proof. All matrices Q of the given form obviously satisfy that $Q^T B$ is symmetric. The following constructive proof shows that this is the only possible form.

Since $U^T B = R$ is symmetric with $U \in O(n)$, the problem is to find all matrices $T = U^T Q \in O(n)$ such that $S = T^T R$ is symmetric and $\det(T) = \det(U)$. Work in a basis of eigenvectors H^* diagonalizing R with its eigenvalues placed in decreasing order $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_n \geq 0$. The following shows that T is diagonal in that basis. Then orthogonality of T imposes value 1 or -1 on the diagonal, the number l of -1 being compatible with $\det(T) = \det(U)$; the final form follows by returning to the original basis and reordering the eigenvectors such that those corresponding to -1 are in the first columns.

The jth column of S is simply the j^{th} column of T multiplied by λ_j. Therefore, we have the following cases:

1. If $\lambda_i = \lambda_j$, then H^* may be chosen such that the corresponding submatrix $T(i : j, i : j) =$ intersection of rows i to j and columns i to j of T is diagonal.

2. If $\lambda_{p+1} = 0$ and $\lambda_p \neq 0$, then S symmetric implies $T(n-p:n,\, 1:p) = 0$. As $T(n-p:n,\, n-p:n)$ is diagonal from case 1, only diagonal elements are nonzero in the last $n-p$ rows of T. Because the rows and columns of T are normalized, $T(1:p,\, n-p:n) = 0$.

3. Consider $i_- \leq p$ and i_+ the smallest index such that $\lambda_{i_+} < \lambda_{i_-}$. Note that

$$\sum_{j=1}^{n} T_{i_-j}^2 = \sum_{j=1}^{n} T_{ji_-}^2 = 1 \text{ (orthogonality) and } \sum_{j=1}^{n} S_{i_-j}^2 = \sum_{j=1}^{n} S_{ji_-}^2 \text{ (symmetry).} \tag{A.1}$$

Start with $i_- = 1$ and assume $\lambda_{i_+} > 0$. Equation (A.1) can be satisfied only if $T_{jk} = T_{kj} = 0\ \forall j \geq i_+$ and $\forall k \in [i,i_+)$; case 1 further implies $T_{jk} = T_{kj} = 0\ \forall j \neq k$ and $\forall k \in [i_-, i_+)$. This argument is repeated by defining the new i_- as being the previous i_+ until $\lambda_{i_+} = 0$ (case 2) or $\lambda_{i_-} = \lambda_n > 0$. This leaves T diagonal. □

Acknowledgments. The authors thank P.-A. Absil for valuable discussions and the anonymous reviewers for helpful comments.

REFERENCES

[1] P. A. Absil, R. Mahony, and R. Sepulchre, *Riemannian geometry of Grassmann manifolds with a view on algorithmic computation*, Acta Appl. Math., 80 (2004), pp. 199–220.

[2] P. A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, Princeton, NJ, 2007.

[3] A. Barg, *Extremal problems of coding theory*, in Coding Theory and Cryptography, H. Niederreiter, ed., Notes of Lectures at the Institute for Mathematical Sciences of the University of Singapore, World Scientific, Singapore, 2002, pp. 1–48.

[4] A. Barg and D.Y. Nogin, *Bounds on packings of spheres in the Grassmann manifold*, IEEE Trans. Inform. Theory, 48 (2002), pp. 2450–2454.

[5] A. K. Bondhus, K. Y. Pettersen, and J. T. Gravdahl, *Leader/follower synchronization of satellite attitude without angular velocity measurements*, in Proceedings of the 44th IEEE Conference on Decision and Control, IEEE Press, Piscataway, NJ, 2005, pp. 7270–7277.

[6] R. W. Brockett, *Dynamical systems that sort lists, diagonalize matrices, and solve linear programming problems*, Linear Algebra Appl., 146 (1991), pp. 79–91.

[7] M. Brookes, *Matrix Reference Manual*, Imperial College London, London, UK, 2005. Available online at http://www.ee.ie.ac.uk/hp/staff/dmb/matrix/intro.html.

[8] F. Bullo and R. Murray (advisor), *Nonlinear Control of Mechanical Systems: A Riemannian Geometry Approach*, Ph.D. thesis, CalTech, Pasadena, CA, 1998.

[9] S. R. Buss and J. P. Fillmore, *Spherical averages and applications to spherical splines and interpolation*, ACM Trans. Graphics, 20 (2001), pp. 95–126.

[10] F. R. K. Chung, *Spectral Graph Theory*, Reg. Conf. Ser. Math. 92, AMS, Providence, RI, 1997.

[11] J. H. Conway, R. H. Hardin, and N. J. A. Sloane, *Packing lines, planes, etc.: Packings in Grassmannian spaces*, Exper. Math., 5 (1996), pp. 139–159.

[12] J. Cortes, S. Martinez, and F. Bullo, *Coordinated deployment of mobile sensing networks with limited-range interactions*, in Proceedings of the 43rd IEEE Conference on Decision and Control, IEEE Press, Piscataway, NJ, 2004, pp. 1944–1949.

[13] T. D. Downs, *Orientation statistics*, Biometrika, 59 (1972), pp. 665–676.

[14] A. Edelman, T. A. Arias, and S. T. Smith, *The geometry of algorithms with orthogonality constraints*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 303–353.

[15] G. A. Galperin, *A concept of the mass center of a system of material points in the constant curvature spaces*, Comm. Math. Phys., 154 (1993), pp. 63–84.

[16] D. Groisser, *Newton's method, zeroes of vector fields, and the Riemannian center of mass*, Adv. Appl. Math., 33 (2004), pp. 95–135.

[17] P. Gruber and F. J. Theis, *Grassmann clustering*, in Proceedings of the 14th European Signal Processing Conference, Florence, Italy, 2006.

[18] U. Helmke and J. B. Moore, *Optimization and Dynamical Systems*, Springer-Verlag, London, 1994.

[19] J. J. Hopfield, *Neural networks and physical systems with emergent collective computational capabilities*, Proc. Nat. Acad. Sci., 79 (1982), pp. 2554–2558.

[20] K. Hueper and J. Manton, *The Karcher mean of points on $SO(n)$*, Talk presented at Cesame (UCL, Belgium), 2004.

[21] M. Hurley, *Chain recurrence, semiflows, and gradients*, J. Dynam. Differential Equations, 7 (1995), pp. 437–456.

[22] A. Jadbabaie, J. Lin, and A. S. Morse, *Coordination of groups of mobile autonomous agents using nearest neighbor rules*, IEEE Trans. Automat. Control, 48 (2003), pp. 988–1001.

[23] E. Justh and P. Krishnaprasad, *A Simple Control Law for UAV Formation Flying*, Tech. report TR 2002-38, ISR, University of Maryland, College Park, MD, 2002.

[24] H. Karcher, *Riemannian center of mass and mollifier smoothing*, Comm. Pure Appl. Math., 30 (1977), pp. 509–541.

[25] T. R. Krogstad and J. T. Gravdahl, *Coordinated attitude control of satellites in formation*, in Group Coordination and Cooperative Control, Lecture Notes in Control and Inform. Sci. 336, Springer-Verlag, Berlin, 2006, pp. 153–170.

[26] Y. Kuramoto, *Self-entertainment of population of coupled nonlinear oscillators*, in Proceedings of the International Symposium on Mathematical Problems in Theoretical Physics, Lecture Notes in Phys. 39, Springer-Verlag, Berlin, 1975, pp. 420–422.

[27] J. R. Lawton and R. W. Beard, *Synchronized multiple spacecraft rotations*, Automatica, 38 (2002), pp. 1359–1364.

[28] N. E. Leonard, D. Paley, F. Lekien, R. Sepulchre, D. Frantantoni, and R. Davis, *Collective motion, sensor networks and ocean sampling*, Proc. IEEE, 95 (2007), pp. 48–74.

[29] A. Machado and I. Salavessa, *Grassmannian manifolds as subsets of Euclidean spaces*, Res. Notes in Math., 131 (1985), pp. 85–102.

[30] K. Mischaikow, H. Smith, and H. R. Thieme, *Asymptotically autonomous semi-flows, chain recurrence and Lyapunov functions*, Trans. AMS, 347 (1995), pp. 1669–1685.

[31] M. Moakher, *Means and averaging in the group of rotations*, SIAM J. Matrix Anal. Appl., 24 (2002), pp. 1–16.

[32] L. Moreau, *Stability of continuous-time distributed consensus algorithms*, in Proceedings of the 43rd IEEE Conference on Decision and Control, IEEE Press, Piscataway, NJ, 2004, pp. 3998–4003.

[33] L. Moreau, *Stability of multi-agent systems with time-dependent communication links*, IEEE Trans. Automat. Control, 50 (2005), pp. 169–182.

[34] S. Nair and N. E. Leonard, *Stabilization of a coordinated network of rotating rigid bodies*, in Proceedings of the 43rd IEEE Conference on Decision and Control, IEEE Press, Piscataway, NJ, 2004, pp. 4690–4695.

[35] R. Olfati-Saber and R. M. Murray, *Consensus problems in networks of agents with switching topology and time-delays*, IEEE Trans. Automat. Control, 49 (2004), pp. 1520–1533.

[36] X. Pennec, *Probabilities and Statistics on Riemannian Manifolds: A Geometric Approach*, Research report 5093, INRIA, Le Chesnay, France, 2004.

[37] A. Sarlette, R. Sepulchre, and N. E. Leonard, *Discrete-time synchronization on the N-torus*, in Proceedings of the 17th Annual International Symposium on Mathematical Theory of Networks and Systems, Kyoto, Japan, 2006, pp. 2408–2414.

[38] A. Sarlette, R. Sepulchre, and N. E. Leonard, *Cooperative attitude synchronization in satellite swarms: A consensus approach*, in Proceedings of the 17th IFAC Symposium on Automatic Control in Aerospace, Toulouse, France, D. Alazard, ed., CD-Rom, 2007.

[39] L. Scardovi, A. Sarlette, and R. Sepulchre, *Synchronization and balancing on the N-torus*, Systems Control Lett., 56 (2007), pp. 335–341.

[40] L. Scardovi and R. Sepulchre, *Collective optimization over average quantities*, in Proceedings of the 45th IEEE Conference on Decision and Control, IEEE Press, Piscataway, NJ, 2006, pp. 3369–3374.

[41] R. Sepulchre, D. Paley, and N. E. Leonard, *Group coordination and cooperative control of steered particles in the plane*, in Group Coordination and Cooperative Control, Lecture Notes in Control and Inform. Sci. 336, Springer-Berlin, Verlag, 2006, pp. 217–232.

[42] R. Sepulchre, D. Paley, and N. E. Leonard, *Stabilization of planar collective motion with all-to-all communication*, IEEE Trans. Automat. Control, 52 (2007), pp. 811–824.

[43] R. Sepulchre, D. Paley, and N. E. Leonard, *Stabilization of planar collective motion with limited communication*, IEEE Trans. Automat. Control, 53 (2008), pp. 706–719.

[44] S. H. Strogatz, *From Kuramoto to Crawford: Exploring the onset of synchronization in populations of coupled nonlinear oscillators*, Phys. D, 143 (2000), pp. 1–20.

[45] S. H. Strogatz, *Sync: The Emerging Science of Spontaneous Order*, Hyperion, New York, 2003.

[46] J. N. Tsitsiklis and D. P. Bertsekas, *Distributed asynchronous optimal routing in data networks*, IEEE Trans. Automat. Control, 31 (1986), pp. 325–332.

[47] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans, *Distributed asynchronous deterministic and stochastic gradient optimization algorithms*, IEEE Trans. Automat. Control, 31 (1986), pp. 803–812.

[48] T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, and O. Shochet, *Novel type of phase transition in a system of self-driven particles*, Phys. Rev. Lett., 75 (1995), pp. 1226–1229.

[49] J. C. Willems, *Lyapunov functions for diagonally dominant systems*, Automatica, 12 (1976), pp. 519–523.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 1, pp. 77–101

© 2009 Society for Industrial and Applied Mathematics

CONSENSUS PROBLEMS WITH DISTRIBUTED DELAYS, WITH APPLICATION TO TRAFFIC FLOW MODELS[*]

WIM MICHIELS[†], CONSTANTIN-IRINEL MORĂRESCU[‡], AND SILVIU-IULIAN NICULESCU[§]

Abstract. This paper focuses on consensus problems for a class of linear systems with distributed delay that are encountered in modeling traffic flow dynamics. In the application problems the distributed delay, whose kernel is a γ-distribution with a gap, represents the human drivers' behavior in the average. The aim of the paper is to give a characterization of the regions in the corresponding delay parameter space, where a consensus is reached for all initial conditions. The structure and properties of the system are fully exploited, which leads to explicit and computationally tractable expressions. As a by-product a stability theory for distributed delay systems with a γ-distribution kernel is developed. Also explicit expressions for the consensus function(al) of time-delay systems with constant and distributed delays that solve a consensus problem are provided. Several illustrative examples complete the presentation.

Key words. consensus problem, time delays, traffic model, stability

AMS subject classifications. 93C23, 93D09, 93C95

DOI. 10.1137/060671425

1. Introduction.

Networks of agents are typically large-scale interconnected systems whose dynamics depend on the topology of the network but also on the individual behaviors of the agents. In this context, agreements and cooperation between agents are needed in order to achieve some common, global objective. Roughly speaking, in a general setting the realization of a *consensus* consists of finding rules and strategies for reaching an agreement regarding some certain quantity of interest depending on the states of all the agents (see, e.g., [20, 23] for a recent survey on the topic).

Consider now a *consensus protocol* for a *multiagent system* with a fixed, directed network topology and a distributed delay in the communication channels (see, e.g., [19, 13] and the references therein). More precisely, let the directed graph

$$\mathcal{G}(V, E, \mathbb{A}) \tag{1.1}$$

[*] Received by the editors October 4, 2006; accepted for publication (in revised form) October 14, 2008; published electronically February 11, 2009. This article presents results of the Research Project IAP P5, funded by the programme on Interuniversity Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Science, Technology and Culture and of OPTEC, the Optimization in Engineering Center of the K. U. Leuven. The work of the first and third authors was partially funded by PAI Tournesol Bilateral Project (France, Flanders): *"Distributed delays in dynamical systems: Analysis and applications"* (2006–2007).

http://www.siam.org/journals/sicon/48-1/67142.html

[†] Department of Computer Science, Katholieke Universiteit Leuven, Celestijnenlaan 200A, 3001 Heverlee, Belgium (Wim.Michiels@cs.kuleuven.be). The author is a postdoctoral fellow of the Fund for Scientific Research-Flanders (Belgium).

[‡] INRIA Rhône Alpes, BIPOP project, Inovalée, 655 avenue de l'Europe, 38330, Montbonnot, France (constantin.morarescu@inrialpes.fr) and Department of Mathematics, University "Politehnica" of Bucharest, Bucharest, Romania. The work of this author was (partially) supported through a European Community Marie Curie Fellowship and in the framework of the CTS, contract HPMT-CT-2001-00278.

[§] Laboratoire de Signaux et Systèmes (L2S, UMR CNRS 8506), CNRS-Supélec, 3, rue Joliot Curie, 91192, Gif-sur-Yvette, France (Silviu.Niculescu@hds.utc.fr).

be characterized by the node set $V = \{1, \dots, p\}$, a set of edges E where $(k, l) \in E$ if and only if $\alpha_{k,l} \neq 0$, and a weighted adjacency matrix $\mathbb{A}$ with zero diagonal entries and nondiagonal entries equal to $\alpha_{k,l}$. Let each node correspond to an agent whose dynamics are described by

$$\dot{v}_k(t) = u_k(t), \qquad k = 1, \dots, p. \tag{1.2}$$

Furthermore, consider the following protocol:

$$u_k(t) = \sum_{(k,l)\in E} \alpha_{k,l} \int_0^{\infty} f(\theta)(v_l(t-\theta) - v_k(t-\theta))\,\mathrm{d}\theta, \qquad k = 1, \dots, p, \tag{1.3}$$

where $f(\cdot)$ denotes some delay kernel, and the notation $\lfloor \cdot \rfloor$ stands for

$$\lfloor l \rfloor = \begin{cases} l, & l = 1, \dots, p, \\ \lfloor l + p \rfloor, & l < 1. \end{cases}$$

We assume that

$$\alpha_{k,l} \geq 0, \qquad k = 1, \dots, p, \ l = 1, \dots, p, \ k \neq l, \tag{1.4}$$

and that the graph $\mathcal{G}$ is *strongly connected* (see, e.g., [19] for the definition).

Let us discuss some of the motivations for introducing such a model. It is well known and well accepted that networks and, more general, interconnected systems are subject to *propagation* and *communication delays*. If such delays are not critical in the perception and the observation of various behaviors, they may become critical if they are used in decision-making, control, or consensus problems. The proposed model enters in this category.

In this context, most of the cases treated and presented in the literature consider only *constant* (piecewise) delays. If such an assumption can be seen as sufficient for some examples, it becomes quite restrictive and conservative for others and among the applications concerned by such an argument, we can cite *traffic dynamics*. Note that the corresponding models are inherently time delayed because of the limited sensing and acting capabilities of drivers against velocity and position variations [8, 12].

The idea of using delays in traffic flow dynamics is not new and, to the best of the authors' knowledge, was pointed out in the 1960s (see, for instance, [2]). According to its origin (see, e.g., [8]), we can classify the delays in the traffic flow dynamics as follows: physiological delays (mainly induced by the human operators), mechanical time delays (time needed for the vehicle's response after some driver's action), and delays in the vehicles' action, to cite only a few (see also [25]).

Without any deep discussions on the modeling of the traffic dynamics, one of the simplest model often discussed in the literature is the (microscopic) car-following model, describing the behavior of multiple vehicles under the influence of a single constant time delay [2, 12, 24]. In general, two spatial configurations are dealt with: the linear and the ring configurations. In what follows, for the sake of brevity, we shall only consider the *ring configuration* when discussing the traffic flow application, but the obtained results can also be applied to the linear configuration (see, e.g., [27, 26] for further discussions concerning these configurations). The linear model of [2] can be written conceptually as follows:

$$\dot{v}_k(t) = \alpha_k(v_{k-1}(t-\tau) - v_k(t-\tau)), \qquad k = 1, \dots, p, \tag{1.5}$$

where p is the number of considered vehicles and $v_0 = v_p$. The left-hand side represents the *acceleration* of the kth vehicle, and the right-hand side expresses the *velocity difference* of consecutive vehicles (see also [24] for a multiple-car model). One of the limitations of the model above is that, in general, humans retain a short-term memory of the past events and this may affect their control decision strategy and such a behavior cannot be described by using pointwise (or discrete) delays in the model. Furthermore, the drivers' perception and interpretations of the stimuli depend on various parameters and are different from one driver to another. As pointed out in [27], a more realistic model should include a *delay distribution* over the time that depicts the human behavior in average. Conceptually, defining the delay distribution represents a challenging problem itself and is far from being solved. In [27], the authors proposed three types of delay distributions: a uniform distribution, a γ-distribution, and a γ-distribution with a gap, where the gap corresponds to the minimum reaction time of the humans with respect to some external signals and/or stimuli. In this paper we shall assume the third type of distribution. Remarks and discussions on its applications to other problems from engineering and biology can be found in [16]. Finally, it is important to point out that *distributed delays* are often encountered in controlling time-delay systems. Typical examples are given by the finite-spectrum assignment problems or the approximation of the derivative actions by its delay-difference counterpart (see, e.g., [9, 15]).

The above discussions lead us to the analysis of the model

$$\dot{v}_k(t) = \sum_{i=1}^{p-1} \alpha_{k,\lfloor k-i \rfloor} \int_0^\infty f(\theta)(v_{\lfloor k-i \rfloor}(t-\theta) - v_k(t-\theta))\,\mathrm{d}\theta, \qquad k = 1, \ldots, p, \tag{1.6}$$

where, as mentioned above, $f(\cdot)$ denotes the delay kernel. It is easy to see that the model (1.6) is nothing else than the protocol (1.3) applied to the agents described by (1.2). Since the delay distribution is assumed to be a γ-distribution with a gap, the kernel f is given by

$$f(\xi) = \begin{cases} 0, & \xi < \tau, \\ \frac{(\xi-\tau)^{n-1} \mathrm{e}^{-\frac{\xi-\tau}{T}}}{T^n (n-1)!}, & \xi \geq \tau, \end{cases} \tag{1.7}$$

where $n \in \mathbb{N}$, $T > 0$, and $\tau \geq 0$. Note that $f(\xi) \geq 0$ for all $\xi \geq 0$ and $\int_0^\infty f(\xi)\, d\xi = 1$. The gap is defined by τ, and the corresponding *average delay* of (1.7) satisfies

$$\tau_m = \int_0^\infty \xi f(\xi)\, d\xi = \tau + nT.$$

As $T \to 0+$, the kernel (1.7) tends to a Dirac impulse centered at $\xi = \tau$ and (1.6) therefore reduces to a system with a pointwise delay τ. As we shall see, the transition to $T = 0$ is smooth from a stability point of view, as the stability determining eigenvalues are continuous with respect to $T \geq 0$.

The aim of the paper is to analyze the general consensus problem (1.2)–(1.3) for a particular delay kernel (γ-distribution with a gap), more precisely, to perform a stability analysis of (1.6)–(1.7) with respect to the parameters (T, τ) and n, which determine the shape of the delay distribution. For a given value of n we will determine regions in the (T, τ) space, such that for all initial conditions a consensus is reached. In the traffic flow application this corresponds to the fact that all cars eventually get the same speed. Note that the corresponding problem for an undirected graph (where $\mathbb{A}$ is symmetric) and a constant time delay was investigated in [19].

The structure of the paper is as follows. In section 2 the stability theory for systems with distributed delay is addressed, with the emphasis on consensus problems. Section 3 is devoted to the derivation of conditions on the pair (T, τ) for which (1.6)–(1.7) realizes a consensus. Illustrative examples are given in section 4. The analysis of other types of delay models for modeling traffic flow dynamics, in particular the so-called optimal velocity models, is briefly commented on in section 5. Some concluding remarks are formulated in section 6.

The following standard notation will be adopted: $\mathbb{C}$ ($\mathbb{C}^+$, $\mathbb{C}^-$) is the set of complex numbers (with strictly positive and strictly negative real parts), and $j = \sqrt{-1}$. For $z \in \mathbb{C}$, $\angle(z) \in (-\pi\ \pi]$, $\Re(z)$, and $\Im(z)$ define the argument, the real part, and the imaginary part of z. $\mathbb{R}$ ($\mathbb{R}^+$, $\mathbb{R}^-$) denotes the set of real numbers (larger than or equal to zero, smaller than or equal to zero). $\mathbb{N}$ is the set of natural numbers, including zero and $\mathbb{Z}$ is the set of integers. The set $\mathcal{C}(\mathcal{I}, \mathbb{C}^p)$, with $\mathcal{I} \subseteq \mathbb{R} \cup \{\pm\infty\}$ and $p \in \mathbb{N}$, is the space of continuous functions from $\mathcal{I}$ to $\mathbb{C}^p$. Finally, the following functions will be used.

Definition 1.1. *For $n \in \mathbb{N}$, let $g_n : \mathbb{R}^+ \to \mathbb{R}^+$ be such that $y = g_n(x)$ is the positive solution of $|y(1 + jy)^n| = x$.*

2. Stability theory for systems with distributed delays.

Motivated by the structure of (1.6)–(1.7) we develop a stability theory for systems with an unbounded distributed delay of the form

$$\dot{x}(t) = A \int_0^{\infty} f(\theta) x(t - \theta)\, d\theta, \tag{2.1}$$

where $x(t) \in \mathbb{C}^{p\times 1}$, $A \in \mathbb{C}^{p\times p}$, and f is given by (1.7). The approach is based on establishing relations with stability properties of general systems with bounded delay of the form

$$\dot{x}(t) = \int_{-\tau}^{0} d\eta(\theta)\, x(t + \theta),\ \ x(t) \in \mathbb{C}^{r\times r}, \tag{2.2}$$

where $\eta(\theta)$, $\theta \in [-\tau, 0]$, is an $r \times r$ matrix whose elements are of bounded variation, because for such systems a well-established stability theory exists [10].

A solution of (2.1) is uniquely determined for an initial condition ϕ, which belongs to the set $\mathcal{F}((-\infty, 0], \mathbb{C}^{p\times 1})$, defined as

$$\mathcal{F}((-\infty, 0], \mathbb{C}^{p\times 1}) := \left\{\phi \in \mathcal{C}((-\infty, 0], \mathbb{C}^{p\times 1}) :\ \|\phi\|_f := \int_{-\infty}^{0} \|f(-\theta)\phi(\theta)\|_2\, d\theta < \infty\right\}$$

and equipped with $\|\cdot\|_f$. Denote by $t \in (-\infty, \infty) \to x(\phi)(t)$ the forward solution of (2.1) with initial condition ϕ. In this way, stability definitions can be formulated in a similar way as for systems with constant delays; see, e.g., [10, 15] for the latter. We say, for instance, that the zero solution of (2.1) is asymptotically stable if and only if

$$\forall\epsilon > 0\ \exists\delta > 0,\quad \forall\phi \in \mathcal{F}((-\infty, 0], \mathbb{C}^{p\times 1})\ \ \|\phi\|_f < \delta \Rightarrow \forall t \geq 0\ \|x(\phi)(t)\|_2 < \epsilon,$$
$$\forall\phi \in \mathcal{F}((-\infty, 0], \mathbb{C}^{p\times 1})\ \lim_{t\to\infty} x(\phi)(t) = 0.$$

The substitution of a sample solution of the form $x(t) = e^{st}X$, with $X \in \mathbb{C}^{p\times 1}$, in (2.1) leads us to the *characteristic equation*

$$\det(sI - AF(s)) = 0,$$

where $F(s)$ is the Laplace transform of f. When f is given by (1.7) the characteristic equation above is rewritten as

$$\det\left(sI - A\frac{e^{-s\tau}}{(1+sT)^n}\right) = 0, \tag{2.3}$$

which can be factored as

$$\prod_{k=1}^{p}\left(s - \frac{\mu_k e^{-s\tau}}{(1+sT)^n}\right) = 0, \tag{2.4}$$

with μ_k, $k = 1, \ldots, p$, the eigenvalues of A. As we shall see, the roots distribution of (2.3)–(2.4) determines the stability properties of (2.1). However, the commonly used arguments, which are based on a spectral decomposition of the solutions (see, for instance, [6, 10]), cannot be directly applied to a system of the form (2.1). A major obstacle is the fact that functions of the form $e^{st}X$, $t \leq 0$, do not belong to the space $\mathcal{F}((-\infty, 0], \mathbb{C}^{p\times 1})$ if $\Re(s) < 1/T$. We shall therefore develop arguments based on a *comparison system.*

Formally, with

$$y(t) = \int_0^\infty f(\theta+\tau)x(t-\theta)\,d\theta = \int_{-\infty}^{t} f(t+\tau-\theta)x(\theta)\,d\theta,$$

we get

$$\left\{\begin{array}{lcl} y'(t) & = & \int_{-\infty}^{t} f'(t+\tau-\theta)x(\theta)\,d\theta, \\ & \vdots & \\ y^{(n-1)} & = & \int_{-\infty}^{t} f^{(n-1)}(t+\tau-\theta)x(\theta)\,d\theta, \\ y^{(n)}(t) & = & f^{(n-1)}(\tau)x(t) + \int_{-\infty}^{t} f^{(n)}(t+\tau-\theta)x(\theta)\,d\theta, \end{array}\right.$$

which leads to

$$\begin{aligned}\left(T\tfrac{d}{dt}+I\right)^n y(t) &= T^n f^{(n-1)}(\tau)x(t) + \int_{-\infty}^{t}\left(T\tfrac{d}{dt}+I\right)^n f(t+\tau-\theta)\,d\theta \\ &= x(t).\end{aligned}$$

We conclude that a solution $x(\phi)(t)$ of (2.1) satisfies

$$\left\{\begin{array}{l} \dot{x}(t) = Ay(t-\tau), \\ \left(T\frac{d}{dt}+I\right)^n y(t) = x(t) \end{array}\right. \tag{2.5}$$

for $t \geq 0$, if (2.5), interpreted as an initial value problem, is accordingly initialized with

$$\left\{\begin{array}{ll} x(0) = \phi(0), & \\ y(\theta) = \int_{-\infty}^{\theta} f(\tau+\theta-\xi)\phi(\xi)\,d\xi, & \theta \in [-\tau, 0], \\ y^{(i)}(0) = \int_{-\infty}^{0} f^{(i)}(\tau-\xi)\phi(\xi)\,d\xi, & i = 1, \ldots, n-1. \end{array}\right. \tag{2.6}$$

Note that the integrals on the right-hand side of (2.6) are defined and bounded because $f^{(i)}(\xi)$, $i = 1, \ldots, n-1$, has the same asymptotic behavior as $f(\xi)$ as $\xi \to \infty$.

When letting $z = [x^T\ y^T\ y'^T \cdots (y^{(n-1)})^T]^T$, the *comparison system* (2.5) can be written in a first-order form as

$$\dot{z}(t) = \bar{A}z(t) + \bar{B}z(t-\tau), \tag{2.7}$$

where

$$\bar{A} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & I & & \\ & & & \ddots & \\ & & & & I \\ \frac{I}{T^n} & -\binom{n}{n}\frac{1}{T^n} & & \cdots & -\binom{n}{1}\frac{1}{T^1} \end{bmatrix} \text{ and } \bar{B} = \begin{bmatrix} 0 & A & 0 & \cdots & 0 \\ 0 & 0 & 0 & & 0 \\ \vdots & & & & \vdots \\ 0 & & \cdots & & 0 \end{bmatrix}.$$

The initial conditions for (2.7) are assumed to belong to the space $\mathcal{C}([-\tau,0],\mathbb{C}^{(n+1)p\times 1})$. The next lemma summarizes the established relation between the solutions of (2.1) and (2.7) and also contains a partial converse.

LEMMA 2.1. *If $x(t)$, $t \in \mathbb{R}$, is a solution of (2.1), then there exists a solution $z(t)$, $t \geq -\tau$, of (2.7) such that $[I\ 0\cdots 0]z(t) = x(t)$ for all $t \geq -\tau$. If (2.7) has a solution of the form Ze^{st}, $t \geq -\tau$, where $\Re(s) \geq 0$ and $Z \in \mathbb{C}^{(n+1)p\times 1} \setminus \{0\}$, then $[I\ 0\cdots 0]Ze^{st}$, $t \in \mathbb{R}$, is a nontrivial solution of (2.1).*

Proof. The first assertion follows from the above construction, and an extension of (2.6) on the interval $[-\tau, 0]$.

To prove the second assertion, we partition Z according to the structure of $\bar{A}$ and $\bar{B}$ as $Z = [X^T\ Y_0^T \cdots Y_{n-1}^T]^T$. Substituting Ze^{st} in (2.7) yields

$$\left(sI - A\frac{e^{-s\tau}}{(1+sT)^n}\right)X = 0, \tag{2.8}$$

$$Y_i = \frac{s^i}{(1+sT)^n}X, \quad i = 0,\ldots,n-1. \tag{2.9}$$

It follows that $Z \neq 0$ if and only if $X \neq 0$. Furthermore, (2.8) implies that Xe^{st} satisfies (2.1) for all $t \in \mathbb{R}$. The function Xe^{st}, $t \in \mathbb{R}$, is a solution since $\Re(s) \geq 0$ and thus Xe^{st}, $t \leq 0$, belongs to $\mathcal{F}((-\infty,0],\mathbb{C}^{p\times 1})$. □

The system (2.7) is of the form (2.2) and corresponds to taking

$$\eta(\theta) = \begin{cases} -\bar{A} - \bar{B}, & \theta = -\tau, \\ -\bar{A}, & \theta \in (-\tau, 0), \\ 0, & \theta = 0. \end{cases}$$

From [10, 6] the zero solution of (2.2) is asymptotically stable if and only if all the roots of its characteristic equation,

$$\det\left(sI - \int_{-\tau}^{0} d\eta(\theta)\, e^{s\theta}\right) = 0,$$

are in $\mathbb{C}^-$. For (2.7) the characteristic equation reduces to

$$\det(sI - \bar{A} - \bar{B}e^{-s\tau}) = 0 \tag{2.10}$$

or, equivalently,

$$\det\left(\begin{bmatrix} sI & -Ae^{-s\tau} \\ -I & (1+sT)^n \end{bmatrix}\right) = 0. \tag{2.11}$$

In what follows the roots of (2.11) are called eigenvalues of (2.5).

Note that (2.11) reduces to (2.3) if $s \neq -1/T$. Combining this result with Lemma 2.1 results in the following proposition.

Proposition 2.2. *The zero solution of* (2.1) *is asymptotically stable if and only if all roots of* (2.3) *are in* $\mathbb{C}^-$.

Next, we derive conditions on the characteristic roots for which the linear system with unbounded distributed delay (2.1) solves a consensus problem. This stability property is defined in the following way.

Definition 2.3. *The system* (2.1) *solves a consensus problem if and only if*

$$\forall \phi \in \mathcal{F}((-\infty, 0], \mathbb{C}^{p\times 1}) \ \ \lim_{t\to\infty} x(\phi)(t) = \chi(\phi)\, E_0,$$

where $\chi(\phi) \in \mathbb{C}$ *and* $E_0 = [1 \cdots 1]^T$. *The function* $\chi : \mathcal{F}((-\infty, 0], \mathbb{C}^{p\times 1}) \to \mathbb{C}$ *is called the consensus functional. The system* (2.1) *solves a nontrivial consensus problem if and only if it solves a consensus problem and the consensus functional is not identically zero.*

We follow the same methodology as for the asymptotic stability condition: we first address a consensus problem for a system with bounded delay, and next we treat (2.1) using Lemma 2.1.

Lemma 2.4. *The system* (2.2) *with initial condition* $\phi \in \mathcal{C}([-\tau, 0], \mathbb{C}^{r\times 1})$ *solves a nontrivial consensus problem*[1] *if and only if all the roots of*

$$\det\left(sI - \int_{-\tau}^{0} d\eta(\theta)\, e^{s\theta}\right) = 0 \tag{2.12}$$

are in the open left half plane, excepting a zero root with multiplicity one, and $E_0 = [1 \cdots 1]^T$ *is the right null vector of*

$$\int_{-\tau}^{0} d\eta(\theta). \tag{2.13}$$

The consensus functional $\chi : \mathcal{C}([-\tau, 0], \mathbb{C}^{r\times 1}) \to \mathbb{C}$ *can be expressed as*

$$\chi(\phi) = \frac{V_0^T \left(x(\phi)(\hat{t}) + \int_{-\tau}^{0} \int_{\hat{t}+\theta}^{\hat{t}} d\eta(\theta)\, x(\phi)(\xi)\, d\xi\right)}{V_0^T \left(I + \int_{-\tau}^{0} \int_{\theta}^{0} d\eta(\theta)\, d\xi\right) E_0}, \tag{2.14}$$

where V_0 *is the left null vector of* (2.13) *and*

$$\hat{t} \geq p\tau - \limsup_{R\to\infty} \frac{\log \max_{|s|=R} \det\left(sI - \int_{-\tau}^{0} d\eta(\theta)\, e^{s\theta}\right)}{R}. \tag{2.15}$$

Proof. The first assertion is a trivial corollary of the spectrum determined growth property of the solutions of (2.2); see, e.g., [10, 6]. So we restrict ourselves to the assertions on the form of the consensus functional.

Let $\mathcal{T}(t)$, $t \geq 0$, be the time-integration operator associated with the solutions of (2.2), i.e.,

$$\mathcal{T}(t)\phi = x_t(\phi),$$

[1] For the system (2.2) the definition is similar to Definition 2.3, with the difference that the initial conditions should be taken from the set $\mathcal{C}([-\tau, 0], \mathbb{C}^{r\times 1})$.

where $x_t(\phi) \in \mathcal{C}([-\tau,0],\mathbb{C}^{r\times 1})$ is defined by $x_t(\phi)(\theta) = x(\phi)(t+\theta)$, $\theta \in [-\tau,0]$. Note that $\mathcal{T}(t)$ is a strongly continuous semigroup. The roots of (2.12), which are the eigenvalues of its infinitesimal generator $\mathcal{A}$, are infinite in number, but countable. Denote these eigenvalues by λ_i, $i \geq 0$, with $\lambda_0 = 0$, and let P_{λ_i} be the spectral projection onto the corresponding generalized eigenspace $\mathcal{M}_{\lambda_i}$. By Theorem 8.4 of [6], one can decompose a solution $x(\phi)(t)$ on an interval $[t_1,t_2]$, where $\hat{t} < t_1 \leq t_2 < \infty$, in the following way:

$$x_t(\phi) = \mathcal{T}(t)\,\phi = \sum_{i=0}^{\infty} \mathcal{T}(t)P_{\lambda_i}\,\phi. \tag{2.16}$$

Because $\mathcal{M}_0 = \{\theta \in [-\tau,0] \mapsto cE_0 : c \in \mathbb{C}\}$ we get

$$x_t(\phi) = c_0E_0 + \sum_{i=1}^{\infty} \mathcal{T}(t)P_{\lambda_i}\phi, \tag{2.17}$$

where $c_0 \in \mathbb{C}$. Since $\Re(\lambda_i) < 0$ if $i > 1$, and consequently $\lim_{t\to\infty} \mathcal{T}(t)P_{\lambda_i}\phi = 0$, we have

$$\chi(\phi) = c_0. \tag{2.18}$$

In section 7 of [11] it is shown how the adjoint of the infinitesimal generator $\mathcal{A}$ of $\mathcal{T}(t)$ can be constructed on the space $\mathcal{C}([0,\tau],\mathbb{C}^{1\times r})$, starting with the bilinear form

$$\langle \psi,\phi\rangle = \psi(0)\phi(0) + \int_{-\tau}^{0}\int_{\theta}^{0} \psi(\xi-\theta)\,d\eta(\theta)\,\phi(\xi)\,d\xi, \tag{2.19}$$

where $\psi \in \mathcal{C}([0,\tau],\mathbb{C}^{1\times r})$ and $\phi \in \mathcal{C}([-\tau,0],\mathbb{C}^{r\times 1})$. By Lemma 7.3.6 of this reference, a left eigenfunction of $\mathcal{A}$ is complementary to all right (generalized) eigenfunctions under the bilinear form (2.19), provided that they correspond to different eigenvalues. Furthermore, the left eigenfunction corresponding to the zero eigenvalue is given by the function $\theta \in [0,\tau] \mapsto V_0^T$. From these properties, the decomposition (2.17), and the fact that $\mathcal{M}_{\lambda_i}$ is invariant under $\mathcal{T}(t)$ we get

$$\langle V_0^T, x_t(\phi)\rangle = c_0\langle V_0^T, E_0\rangle + \sum_{i=1}^{\infty} \langle V_0^T, T(t)P_{\lambda_i}\,\phi\rangle = c_0\langle V_0^T, E_0\rangle. \tag{2.20}$$

Expressions (2.18)–(2.20) imply (2.14). □

Remark 2.5. Intuitively it is expected that the value of the consensus functional be some kind of average of the *initial function* over the interval $[-\tau,0]$ and over the agents. Expression (2.14) corresponds to an average of the *state at time* $\hat{t}$, with $\hat{t}$ not necessarily equal to zero. To illustrate the role of condition (2.15) we consider the system

$$\left\{\begin{array}{rcl} \dot{x}_1(t) & = & -x_1(t) + \frac{x_1(t)+x_2(t-\tau)}{2}, \\ \dot{x}_2(t) & = & -x_2(t) + \frac{x_1(t)+x_2(t)}{2}. \end{array}\right. \tag{2.21}$$

From the characteristic equation

$$s^2 + s + \frac{1}{4}e^{-s\tau} = 0 \tag{2.22}$$

it can be seen that it solves a consensus problem for any $\tau > 0$. Because the highest power of $e^{-s\tau}$ in (2.22) is equal to one, condition (2.15) becomes $\hat{t} \geq \tau$, and the consensus functional satisfies (with $\hat{t} = \tau$)

$$V(\phi) = \frac{x_1(\phi)(\tau) + x_2(\phi)(\tau) + \frac{1}{2}\int_0^\tau x_2(\phi)(\xi)\,d\xi}{2 + \frac{1}{2}\tau}.$$

The underlying reason why $\hat{t}$ cannot be chosen equal to zero is that for $t \geq 0$ the solution of (2.21) with initial condition $\phi = [\phi_1^T \ \phi_2^T]^T$ is determined only by

$$\phi_1(0), \quad \phi_2(\theta), \ \theta \in [-\tau, 0], \tag{2.23}$$

while the function segment $\phi_1(\theta)$, $\theta \in [-\tau, 0)$, has no influence on the future behavior and on the value of the consensus functional. By taking $\hat{t} \geq \tau$, i.e., by considering the state at time $\hat{t} \geq \tau$ which only depends on (2.23), the irrelevant part of the initial condition is ignored.

This phenomenon is strongly related to the presence of so-called *small solutions*, that is, solutions that vanish in a finite time (in the example the small solutions satisfy $\phi_2(\theta) = 0$, $\theta \in [-\tau, 0]$, and $\phi_1(0) = 0$). Condition (2.15) ensures that the contributions from such small solutions have disappeared in the solution under consideration. For more details on small solutions we refer the reader to [6, Chapter V].

Remark 2.6. The consensus functional proposed in Theorem 2.1 of [3] is a special case of the general functional (2.14), for compartmental time-delay systems.

COROLLARY 2.7. *If the system*

$$\dot{x}(t) = A_0 x(t - \tau), \qquad x(t) \in \mathbb{C}^{r\times 1}, \tag{2.24}$$

with initial condition $\phi \in \mathcal{C}\left([-\tau, 0], \mathbb{C}^{n\times 1}\right)$ *solves a nontrivial consensus problem, then the consensus functional satisfies*

$$\chi(\phi) = \frac{V_0^T \phi(0)}{V_0^T E_0}. \tag{2.25}$$

Proof. System (2.24) is a special case of (2.2) and can be obtained by taking

$$\eta(\theta) = \begin{cases} -A_1, & \theta = -\tau, \\ 0, & \theta \in (-\tau, 0]. \end{cases}$$

It follows that

$$V_0^T A_1 = 0. \tag{2.26}$$

Furthermore, integrating (2.2),

$$x_{\hat{t}}(\phi) = \phi(0) + \int_0^{\hat{t}} A_1 x(\theta - \tau)\,d\theta$$

makes clear that

$$V_0^T x_{\hat{t}}(\phi) = V_0^T \phi(0). \tag{2.27}$$

Taking into account (2.26) and (2.27), expression (2.14) simplifies to (2.25). □

Proposition 2.8. *The system* (2.1) *solves a nontrivial consensus problem if and only if all roots of* (2.3) *are in the open left half plane, excepting a root at zero with multiplicity one, and* $AE_0 = 0$*, with* $E_0 = [1 \dots 1]^T$*. The corresponding consensus functional* $V : \mathcal{F}((-\infty, 0], \mathbb{C}^{p\times 1}) \to \mathbb{C}$ *satisfies*

$$\chi(\phi) = \frac{V_0^T \phi(0)}{V_0^T E_0}, \tag{2.28}$$

where V_0 *is the left eigenvector of* A *corresponding to the zero eigenvalue.*

Proof. We split the proof of the first assertion into two parts.

($\Rightarrow$) We give a proof by contradiction, which allows to exclude all other possibilities. If (2.3) has a root in $\mathbb{C}^+$ or in $j\mathbb{R} \setminus \{0\}$, then the comparison system (2.7) has a corresponding exponential solution, and by virtue of Lemma 2.1, so does (2.1). This contradicts the fact that the latter solves a consensus problem. The multiplicity of zero as a root of (2.3) is equal to its multiplicity as an eigenvalue of A. If $AE = 0$ and $E \neq E_0$, then $x(t) = E$ is a solution of (2.1), which also leads to a contradiction. Next, we consider the case where zero is a multiple eigenvalue of A, yet with only one eigenvector E_0. Then there exists a generalized eigenvector H_0 such that $AH_0 = E_0$. Consequently, $x(t) = H_0 + tE_0$ is a solution of (2.1) (note that its restriction to $t \leq 0$ belongs to $\mathcal{F}((-\infty, 0], \mathbb{C}^{p\times 1})$) and we arrive again at a contradiction. Finally, if all roots of (2.3) are in $\mathbb{C}^-$, then the zero solution of (1.6) is asymptotically stable (Proposition 2.2), hence it does not solve a nontrivial consensus problem.

($\Leftarrow$) Following Lemma 2.1 a solution $x(\phi)(t)$ of (2.1), restricted to $t \geq \tau$, also appears as a component of a corresponding solution of the comparison system (2.7), which we call $z(\bar{\phi})$ in what follows. The left and right eigenvectors of $\bar{A} + \bar{B}$ corresponding to the zero eigenvalue are given by

$$\bar{V}_0 = [V_0^T \; 0 \cdots 0]^T \text{ and } \bar{E}_0 = [E_0^T \; -E_0^T \; 0 \cdots 0]^T.$$

Note that $\bar{V}_0^T \bar{A} = V_0^T \bar{B} = 0$. Given the condition on the roots of (2.3), one proves, using the same arguments as in the proof of Lemma 2.4 (based on spectral decomposition), that

$$\lim_{t\to\infty} z(\bar{\phi})(t) = \frac{\bar{V}_0^T \left(z_{\hat{t}}(\bar{\phi})(0) + \bar{B} \int_{\hat{t}-\tau}^{\hat{t}} z(\phi)(\theta)\, d\theta \right)}{\bar{V}_0^T (I + \tau \bar{B}) \bar{E}_0} \bar{E}_0 = \frac{V_0^T \phi(0)}{V_0^T E_0} \bar{E}_0, \tag{2.29}$$

where

$$\hat{t} \geq p(n+1)\tau - \limsup_{R\to\infty} \frac{\log F(R)}{R}, \qquad F(R) = \max_{|s|=R} \det\left(sI - \bar{A} - \bar{B}e^{-s\tau}\right).$$

It follows that

$$\lim_{t\to\infty} x(\phi)(t) = \frac{V_0^T \phi(0)}{V_0^T E_0} E_0, \tag{2.30}$$

implying that (2.7) solves a consensus problem.

The assertion on the consensus function follows from (2.30). □

Remark 2.9. Expressions (2.25) and (2.28) also follow from a simple geometric argument. As in both cases $V_0^T \dot{x}(t) = 0$ a solution $x(\phi)(t)$ is constrained to the plane

$V_0^T x = V_0^T \phi(0)$ for all $t \geq 0$. Furthermore, a constant stationary solution must be a multiple of E_0. Thus $x^*(\phi) = \lim_{t\to\infty} x(\phi)(t)$ satisfies the equations

$$\begin{cases} V_0^T x^*(\phi) = V_0^T \phi(0), \\ x^*(\phi) = \chi(\phi) E_0, \end{cases}$$

which can be interpreted as the intersection of the plane through $\phi(0)$ and perpendicular to V_0 with a line with slope E_0. A similar argument was used in section X of [19].

3. Conditions for the realization of a consensus. We perform a stability analysis of the system (1.6)–(1.7) in the (T, τ) parameter space. In particular, we give necessary and sufficient conditions such that it solves a consensus problem.

The system (1.6)–(1.7) can be written in the form (2.1), yet has some special properties due to the induced structure of A, which we outline first. Next, we make an analysis of an auxiliary scalar equation and, finally, we present the main results.

3.1. Properties. The system (1.6)–(1.7) is of the form (2.1), where $A = [a_{k,l}]$ is defined as

$$a_{k,l} = \begin{cases} \alpha_{k,l}, & k \neq l, \\ -\sum_{i=1,\, i\neq k}^{p} \alpha_{k,i}, & k = l. \end{cases} \tag{3.1}$$

Note that in the context of multiagent systems $-A$ is typically called the graph Laplacian of (1.1). By construction A has the following property.

PROPERTY 3.1. *All eigenvalues of A, defined by* (3.1), *are in $\mathbb{C}^-$, excepting a zero eigenvalue with multiplicity one.*

Proof. A is a Metzler matrix with zero row sums. Furthermore, the graph (1.1) is strongly connected. Under these conditions the statement of the proposition follows from Theorems 1 and 2 of [19]. □

Note that zero also appears as a root of (2.3) and (3.1), whatever the values of T, τ, and n. If all other roots are in $\mathbb{C}^-$, we have from Proposition 2.8 that the system (1.6)–(1.7) solves a (nontrivial) consensus problem with delay. In the car-following application the consensus variables are the speed of the vehicles. This means that, whatever the initial values, the speed of the vehicles will eventually converge to a common value (which depends on the initial values). In what follows we shall use the following terminology to characterize parameter values in the (T, τ) space for which a consensus is reached.

DEFINITION 3.2. *The consensus region of* (1.6)–(1.7) *in the* (T, τ) *parameter space is the set of parameters* (T, τ) *for which the system* (1.6)–(1.7) *solves a consensus problem.*

3.2. Analysis of an auxiliary function. Motivated by Property 3.1 and the factorization of the characteristic equation (2.3) as (2.4) we are led to study the zeros of the function

$$\xi(s; T, \tau) := s(1 + sT)^n \mathrm{e}^{s\tau} - \mu, \qquad \mu \in \mathbb{C}^-, \tag{3.2}$$

as a function of the parameters T and τ. Note that the zeros of (3.2) are in $\mathbb{C}^-$ if and only if the roots of

$$s - \frac{\mu e^{-s\tau}}{(1 + sT)^n} = 0$$

are in $\mathbb{C}^-$. We need the following lemmas.

LEMMA 3.3. *In the parameter domain a change of the number of zeros ξ in the closed right half plane is invariably associated with zeros crossing the imaginary axis.*

Proof. The proof follows from the continuous dependence of the individual zeros with respect to the parameters and the fact that the zeros of ξ in the closed right half plane satisfy

$$|s| \leq |\mu| \frac{|e^{-s\tau}|}{|1+sT|^n} \leq |\mu|,$$

which excludes roots coming from the point at infinity. □

LEMMA 3.4. *If the function* (3.2) *has a zero on the imaginary axis, then the multiplicity of this zero is equal to one. Furthermore, an increase of τ leads to a crossing towards $\mathbb{C}^+$. If $\tau = 0$, then also an increase of T leads to a crossing towards $\mathbb{C}^+$.*

Proof. The first assertion is due to the fact that

$$\frac{\partial \xi}{\partial s}(j\omega; T, \tau) = (1+j\omega T)^{n-1} e^{j\omega\tau} \left(1 + j\omega T + j\omega n + j\omega\tau(1+j\omega T)\right)$$

is nonzero for all $\omega \in \mathbb{R}$.

Next, let $j\bar{\omega}$, $\bar{\omega} \in \mathbb{R}$, be an isolated zero of (3.2) for $(T,\tau) = (\bar{T}, \bar{\tau})$. It is clear that $\bar{\omega} \neq 0$. Since the zeros of (3.2) behave continuously with respect to the parameters T and τ, there exists a function

$$r : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}, \qquad (T,\tau) \mapsto r(T,\tau),$$

satisfying $r(\bar{T}, \bar{\tau}) = j\bar{\omega}$ and

$$\xi(r(T,\tau); T, \tau) = 0. \tag{3.3}$$

Differentiating (3.3) at $(\bar{T}, \bar{\tau})$ yields

$$\frac{\partial r}{\partial \tau}(\bar{T}, \bar{\tau}) = -\frac{\frac{\partial \xi}{\partial \tau}(j\bar{\omega}; \bar{T}, \bar{\tau})}{\frac{\partial \xi}{\partial s}(j\bar{\omega}; \bar{T}, \bar{\tau})}, \qquad \frac{\partial r}{\partial T}(\bar{T}, \bar{\tau}) = -\frac{\frac{\partial \xi}{\partial T}(j\bar{\omega}; \bar{T}, \bar{\tau})}{\frac{\partial \xi}{\partial s}(j\bar{\omega}; \bar{T}, \bar{\tau})},$$

from which one obtains

$$\Re\left(\left(\tfrac{\partial r}{\partial \tau}(\bar{T}, \bar{\tau})\right)^{-1}\right) = \tfrac{1}{\bar{\omega}^2} + \tfrac{n\bar{T}}{1+\bar{\omega}^2\bar{T}^2}, \qquad \Re\left(\left(\tfrac{\partial r}{\partial T}(\bar{T}, \bar{\tau})\right)^{-1}\right) = \tfrac{1}{\bar{\omega}^2} - \tfrac{\bar{T}\bar{\tau}}{n}.$$

The first expression is strictly positive; the second is if $\bar{\tau} = 0$. This implies the second assertion. □

We adopt a two-stage approach, similar to the one proposed by [17]. First we characterize the zeros distribution of (3.2) as a function of T, under the condition $\tau = 0$.

PROPOSITION 3.5. *If μ is real and $n = 1$, then the zeros of $\xi(s; T, 0)$ are in $\mathbb{C}^-$ for all $T \geq 0$. Otherwise, the zeros of $\xi(s; T, 0)$ are in $\mathbb{C}^-$ if and only if $T \in [0, T_\mu)$, where*

$$T_\mu = \frac{\tan\left(\dfrac{|\angle(\mu)| - \frac{\pi}{2}}{n}\right)}{|\mu| \left[\cos\left(\dfrac{|\angle(\mu)| - \frac{\pi}{2}}{n}\right)\right]^n}. \tag{3.4}$$

Proof. Assume that $\xi(s;T,0)$ has a zero on the imaginary axis for some value of T. Then there exists a corresponding frequency $\omega>0$ such that either

$$\xi(j\omega/T;T,0)=0 \text{ or } \xi(-j\omega/T;T,0)=0,$$

which is equivalent to

$$T=\frac{\pm j\omega(1\pm j\omega)^n}{\mu}. \tag{3.5}$$

If $n=1$ and μ is real, then the right-hand side of (3.5) cannot be real, whatever the value of ω. Hence, $\xi(s;T,0)$ cannot have zeros on the imaginary axis. Combining this fact with the continuity of the zeros with respect to T and the assumption $\mu\in\mathbb{C}^-$, i.e., all zeros are in $\mathbb{C}^-$ for $T=0$, yields the first statement of the proposition.

If $n>1$ or $\Im(\mu)\neq 0$, then there always exist pairs (T,ω), $T>0$, which satisfy (3.5), that is, $\xi(s;T,0)$ has zero $+j\omega/T$ or $-j\omega/T$. Because, by Lemma 3.4, the corresponding crossing direction of the imaginary axis is towards $\mathbb{C}^+$ as T is increased, and because $\mu\in\mathbb{C}^-$, all zeros of $\xi(s;T,0)$ are in $\mathbb{C}^-$ if and only if $T\in[0,T_\mu)$, where

$$T_\mu=\min\left\{T>0:(T,\omega)\text{ satisfies (3.5) for some }\omega>0\right\}.$$

Since the functions $\omega>0\rightarrow|\pm j\omega(1\pm j\omega)^n|/|\mu|$ are strictly increasing, T_μ is determined by the *first* intersection of one of the two curves

$$\omega>0\rightarrow\frac{\pm j\omega(1\pm j\omega)^n}{\mu} \tag{3.6}$$

with the positive real axis, as ω is increased from zero. In what follows we distinguish between two cases.

Case 1. $\Im(\mu)\geq 0$. The first intersection of (3.6) with the positive real axis is due to the curve corresponding to the plus sign. It is characterized by $\omega=\bar{\omega}$, satisfying

$$\angle(j\bar{\omega})+\angle((1+j\bar{\omega})^n)-\angle(\mu)=0.$$

It follows that

$$\bar{\omega}=\tan\left(\angle(1+j\bar{\omega})\right)=\tan\left(\frac{\angle(\mu)-\pi/2}{n}\right) \tag{3.7}$$

and

$$(1+j\bar{\omega})^n=(1+\bar{\omega}^2)^{n/2}e^{j(\angle\mu-\pi/2)}=\frac{e^{j(\angle(\mu)-\pi/2)}}{\left[\cos\left(\frac{\angle(\mu)-\pi/2}{n}\right)\right]^n}. \tag{3.8}$$

Expression (3.4) is obtained when substituting (3.7) and (3.8) in

$$T_\mu=\frac{j\bar{\omega}(1+j\bar{\omega})^n}{\mu}.$$

Case 2. $\Im(\mu)<0$. The first intersection of (3.6) with the positive real axis is due to the curve corresponding to the minus sign and characterized by $\omega=\hat{\omega}$, where

$$\angle(-j\hat{\omega})+\angle((1-j\hat{\omega})^n)+|\angle(\mu)|=0.$$

One can proceed as in the former case. □

Second, we fix T and characterize the zeros distribution of (3.2) as a function of the delay parameter τ. We make use of the functions g_n described by Definition 1.1.

PROPOSITION 3.6. *The function $\xi(s;T,\tau)$, with T fixed, has a zero on the imaginary axis for some delay value τ if and only if $\tau \in \mathcal{T}_\mu(T)$, where*[2]

$$\mathcal{T}_\mu(T) = \left\{ \tau \geq 0 \mid \tau = \frac{\pm\angle(\mu) - \angle(j\omega(1+j\omega)^n) + 2\pi l}{\omega/T}, \quad l \in \mathbb{Z},\ \omega = g_n(T|\mu|) \right\}. \tag{3.9}$$

Furthermore, all zeros are in $\mathbb{C}^-$ if and only if the zeros of $\xi(s;T,0)$ are in $\mathbb{C}^-$ and $\tau \in [0, \tau_\mu(T))$, where

$$\tau_\mu(T) = \frac{|\angle(\mu)| - \angle(j\omega(1+j\omega)^n)}{\omega/T}, \quad \omega = g_n(T|\mu|). \tag{3.10}$$

Proof. The function $\xi(s;T,\tau)$ has a zero $j\omega/T$ or $-j\omega/T$, $\omega > 0$, if and only if

$$j\omega(1+j\omega)^n = T\mu e^{-j\omega\tau/T} \quad \text{or} \quad -j\omega(1-j\omega)^n = T\mu e^{j\omega\tau/T}.$$

Equating modulus and phase of left- and right-hand sides leads to (3.9).

From the second assertion of Lemma 3.4 it follows that $\xi(s;T,\tau)$ has its zeros in the open left half plane if and only if $\xi(s;T,0)$ does so and $\tau \in [0, \tau_m(T))$, where

$$\tau_m(T) = \min_{\tau > 0} \mathcal{T}_\mu(T).$$

It remains to prove that $\tau_m(T) = \tau_\mu(T)$, with $\tau_\mu(T)$ defined by (3.10), for all T such that $\xi(s;T,0)$ has its zeros in $\mathbb{C}^-$ (described by Proposition 3.5).

For sufficiently small T, and thus sufficiently small $\omega(T) = g_n(T|\mu|)$, we have

$$0 < |\angle(\mu)| - \angle(j\omega(T)(1+j\omega(T))^n) \leq \pi/2 \tag{3.11}$$

and

$$\begin{aligned} &-|\angle(\mu)| - \angle(j\omega(T)(1+j\omega(T))^n) + 2\pi \\ &\quad = |\angle(\mu)| - \angle(j\omega(T)(1+j\omega(T))^n) + 2(\pi - |\angle(\mu)|) \\ &\quad \geq |\angle(\mu)| - \angle(j\omega(T)(1+j\omega(T))^n), \end{aligned} \tag{3.12}$$

which imply

$$\tau_m(T) = \tau_\mu(T). \tag{3.13}$$

As the function $T \rightarrow \omega(T)$ is monotonically increasing, (3.11)–(3.12), and consequently (3.13), hold either for all $T \geq 0$ or for values of T which belong to a finite number of intervals. In the latter case, one of these intervals is given by $[0, T_m)$, with T_m satisfying

$$|\angle(\mu)| - \angle(j\omega(T_m)(1+j\omega(T_m))^n) = 0.$$

[2] The right-hand side of expression (3.9) is not defined for $T = 0$, which implies $\omega = g_n(T|\mu|) = 0$. In that case one should interpret $g_n(T|\mu|)/T$ as $\lim_{T\to 0+} g_n(T|\mu|)/T = |\mu|$ and $\angle(j\omega(1+j\omega)^n)$ as $\lim_{\omega\to 0+} \angle(j\omega(1+j\omega)^n) = \pi/2$.

It follows that $0 \in \mathcal{T}_\mu(T_m)$, which implies on its turn that T_μ exists and $T_m \geq T_\mu$. We conclude that $\tau_\mu(T) = \tau_m(T)$ for $T \in [0, T_\mu)$. □

Remark 3.7. In expression (3.9), the plus sign of $\pm\angle\mu$ refers to zeros on the positive imaginary axis, the minus sign to zeros on the negative imaginary axis. If $\Im(\mu) \neq 0$, the corresponding values of τ are in general different.

Finally, combining Propositions 3.5 and 3.6 yields the following proposition.

PROPOSITION 3.8. *If μ is real and $n = 1$, then the zeros of* (3.2) *are in $\mathbb{C}^-$ if and only if $T \in [0, \infty)$ and $\tau \in [0, \tau_\mu(T))$. Otherwise, the zeros are in $\mathbb{C}^-$ if and only if $T \in [0, T_\mu)$ and $\tau \in [0, \tau_\mu(T))$.*

3.3. Main results. Taking into account the factorization of (2.3) as (2.4), Property 3.1, and Proposition 3.8, we obtain the following characterization of the consensus region (cf. Definition 3.2) of the system (1.6)–(1.7) in the (T, τ) space.

THEOREM 3.9. *If $n = 1$ and all eigenvalues of A, defined by* (3.1), *are real, then the consensus region of* (1.6)–(1.7) *in the (T, τ) plane is unbounded and characterized by*

$$T \in [0, \infty), \quad \tau \in [0, \tau^*(T)),$$

where

$$\tau^*(T) = \min_{k=1,\ldots,p,\, \mu_k \neq 0} \frac{|\angle(\mu_k)| - \angle(j\omega_k(T)(1 + j\omega_k(T))^n)}{\frac{\omega_k(T)}{T}} \tag{3.14}$$

and $\omega_k(T) = g_n(T|\mu_k|)$. Otherwise, the consensus region is bounded and characterized by

$$T \in [0, T^*), \quad \tau \in [0, \tau^*(T)),$$

where

$$T^* = \min_{k=1,\ldots,p,\, \Im(\mu_k)>0} \frac{\tan\left(\dfrac{\angle(\mu_k) - \frac{\pi}{2}}{n}\right)}{|\mu_k|\left[\cos\left(\dfrac{\angle(\mu_k) - \frac{\pi}{2}}{n}\right)\right]^n} \tag{3.15}$$

and $\tau^*(T)$ is given by (3.14).

Based on this result the consensus region of (1.6)–(1.7) can be computed fully *automatically.* For large p the overall computational complexity is determined by the computation of the eigenvalues of the $p \times p$ matrix A.

Theorem 3.9 does not make assumptions on the multiplicity of the eigenvalues of A and is generally applicable. If A has eigenvalues with multiplicity larger than one, then the stability study of (1.6)–(1.7) is even facilitated as not all factors in (2.4) are different. The following proposition clarifies the connection between multiple eigenvalues of A and multiple eigenvalues of the comparison system (2.7) of (1.6)–(1.7).

PROPOSITION 3.10. *Let $\hat{\mu}$ be a nonzero eigenvalue of A with multiplicity m_1 and a corresponding eigenspace of dimension m_2. Then the roots of*

$$s(1 + sT)^n e^{s\tau} - \hat{\mu} = 0 \tag{3.16}$$

with multiplicity m_3 are eigenvalues of the comparison system (2.7) *with multiplicity $m_1 m_3$ and an eigenspace of dimension m_2. Furthermore, if $m_3 = 1$, then these roots smoothly depend on the parameters T and τ.*

Proof. Let $\hat{s}$ be a zero of (3.16) with multiplicity m_3. The factorization (2.4) implies that $\hat{s}$ is an eigenvalue of (2.7) with multiplicity $m_1 m_3$. Following [10], the corresponding eigenfunctions have the form $e^{\hat{s}\theta}Z$, $\theta \in [-\tau, 0]$, where $Z \in \mathbb{R}^{p(n+1)\times 1}$ satisfies

$$(\hat{s}I - \bar{A} - \bar{B}e^{-\hat{s}\tau})Z = 0. \tag{3.17}$$

With Z partitioned according to the structure of $\bar{A}$ and $\bar{B}$ as $Z = [X^T\ Y_0^T \cdots Y_{n-1}^T]^T$, writing out (3.17) yields

$$\left(\hat{s}I - A\frac{e^{-\hat{s}\tau}}{(1+\hat{s}T)^n}\right)X = 0, \quad Y_i = \frac{\hat{s}^i}{(1+\hat{s}T)^n}X, \quad i = 0, \dots, n-1.$$

Since

$$\left(\hat{s}I - A\frac{e^{-\hat{s}\tau}}{(1+\hat{s}T)^n}\right)X = \frac{e^{-\hat{s}\tau}}{(1+\hat{s}T)^n}\left(\hat{\mu}I - A\right)X,$$

X must be an eigenvector of A corresponding $\hat{\mu}$. Hence, the dimension of the eigenspace of $\hat{s}$ is equal to m_2.

If $m_3 = 1$, then $\hat{s}$ is an *isolated* root of (3.16), and the last statement can be proven using the arguments spelled out in the proof of Lemma 3.4. □

Remark 3.11. If $m_1 > m_2$, then the roots of (3.16) with multiplicity one (this is, for instance, always the case for roots on the imaginary axis; see Proposition 3.5) are multiple, nonsemisimple eigenvalues of (2.7), yet they smoothly depend on the parameters T and τ. Small changes of T and τ do not lead to a splitting of these multiple eigenvalues.

The next proposition addresses a scaling property of the consensus region.

PROPOSITION 3.12. *If the matrix A is scaled with a factor $\epsilon > 0$, then the consensus region of* (1.6)–(1.7) *in the (T, τ) plane is scaled with a factor ϵ^{-1} in both directions.*

Proof. The proof follows from the scaling property

$$\det\left(sI - \epsilon A\frac{e^{-s\tau}}{(1+Ts)^n}\right) = \epsilon^p \det\left(\bar{s}I - A\frac{e^{-\bar{s}(\epsilon\tau)}}{(1+(\epsilon T)\bar{s})^n}\right), \quad \bar{s} = s/\epsilon. \qquad \square$$

Remark 3.13. Proposition 3.12 implies an inherent trade-off between the rate with which the undelayed system ($\tau = T = 0$) reaches a consensus (determined by the rightmost nonzero eigenvalue of A), and the robustness of this stability property with respect to delays. Such an observation was already made in [19], where the case of a symmetric matrix A and a pointwise delay was dealt with.

In the remainder of this section, we refine Theorem 3.9 to two special cases where exploiting the additional structure leads to a *simpler* characterization of the consensus region, and also allows an *analytical* expression for the solutions corresponding to an onset of instability. The proof of the resulting propositions can be found in the appendix.

The following result corresponds to the situation where all cars/drivers have an identical behavior and the reaction of a driver is determined by the preceding car only.

PROPOSITION 3.14. *Consider the system* (1.6)–(1.7)*, where*

$$\alpha_{k,l} = \begin{cases} \alpha > 0, & \lfloor k - l \rfloor = 1, \\ 0 & \textit{otherwise.} \end{cases} \tag{3.18}$$

If $n = 1$ and $p = 2$, then the consensus region in the (T, τ) plane is unbounded and characterized by

$$T \in [0, \infty), \quad \tau \in [0, \tau^*(T)),$$

where

$$\tau^*(T) = \frac{\frac{\pi}{p} - n \arctan(\omega(T))}{\frac{\omega(T)}{T}}, \quad \omega(T) = g_n\left(2\alpha T \sin\left(\frac{\pi}{p}\right)\right). \tag{3.19}$$

Otherwise, the consensus region is bounded and characterized by

$$T \in [0, T^*), \quad \tau \in [0, \tau^*(T)),$$

where

$$T^* = \frac{\tan\left(\frac{\pi}{pn}\right)}{2\alpha \sin\left(\frac{\pi}{p}\right)\left(\cos\left(\frac{\pi}{pn}\right)\right)^n} \tag{3.20}$$

and $\tau^*(T)$ is given by (3.19).

For $\tau = \tau^*(T)$ the stationary solutions are backwards traveling waves:

$$\begin{bmatrix} v_1^s(t) \\ v_2^s(t) \\ \vdots \\ v_p^s(t) \end{bmatrix} = C_1 \begin{bmatrix} \cos\left(\frac{\omega(T)}{T}t + \varphi\right) \\ \cos\left(\frac{\omega(T)}{T}t + \varphi - \frac{2\pi}{p}\right) \\ \vdots \\ \cos\left(\frac{\omega(T)}{T}t + \varphi - \frac{2\pi(p-1)}{p}\right) \end{bmatrix} + C_2 \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \tag{3.21}$$

where $\omega(T)$ is defined in (3.19) *and the constants C_1, C_2, and ϕ depend on the initial conditions.*

Second, we consider the case where (1.6)–(1.7) is of the form (2.1), with the matrix A symmetric. Although this is not a realistic assumption from the car-*following* application point of view, it makes sense in the context of consensus algorithms for multiagent systems. The symmetry of A there corresponds to an *undirected* network topology.

PROPOSITION 3.15. *Consider the system* (1.6)–(1.7) *with A symmetric. If $n = 1$, then the consensus region of* (1.6)–(1.7) *in the (T, τ) plane is unbounded and characterized by*

$$T \in [0, \infty), \quad \tau \in [0, \tau^*(T)),$$

where

$$\tau^*(T) = \frac{\frac{\pi}{2} - n \arctan(\omega(T))}{\frac{\omega(T)}{T}}, \quad \omega(T) = g_n(T|\lambda_{\max}(A)|). \tag{3.22}$$

Otherwise, the consensus region is bounded and characterized by

$$T \in [0, T^*), \quad \tau \in [0, \tau^*(T)),$$

where

$$T^* = \frac{\tan\left(\frac{\pi}{2n}\right)}{|\lambda_{\max}(A)|\left[\cos\left(\frac{\pi}{2n}\right)\right]^n} \tag{3.23}$$

and $\tau^*(T)$ *is given by* (3.22).

If, in addition,

$$\alpha_{k,l} = \begin{cases} \alpha > 0, & \lfloor k-l \rfloor = 1 \textit{ or } \lfloor l-k \rfloor = 1, \\ 0 & \textit{otherwise,} \end{cases} \tag{3.24}$$

then

$$\lambda_{\max}(A) = \begin{cases} -4\alpha \quad (\textit{multiplicity } 1), & p \ \textit{even,} \\ -2\alpha\left(1+\cos\left(\frac{\pi}{p}\right)\right) \quad (\textit{multiplicity } 2), & p \ \textit{odd.} \end{cases}$$

The stationary solutions for $\tau = \tau^*(T)$ *take the form*

$$\begin{bmatrix} v_1^s(t) \\ \vdots \\ v_{p-1}^s(t) \\ v_p^s(t) \end{bmatrix} = C_1 \begin{bmatrix} (-1)^{p-1} \\ \vdots \\ (-1) \\ 1 \end{bmatrix} \cos\left(\frac{\omega(T)}{T}t + \varphi_1\right) + C_2 \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix} \tag{3.25}$$

if p is even and

$$\begin{aligned} \begin{bmatrix} v_1^s(t) \\ \vdots \\ v_{p-1}^s(t) \\ v_p^s(t) \end{bmatrix} &= C_3 \begin{bmatrix} (-1)^{p-1}\cos\left(\frac{\pi(p-1)}{p}\right) \\ \vdots \\ (-1)\cos\left(\frac{\pi.1}{p}\right) \\ 1 \end{bmatrix} \cos\left(\frac{\omega(T)}{T}t + \varphi_2\right) \\ &\quad + C_4 \begin{bmatrix} (-1)^{p-1}\sin\left(\frac{\pi(p-1)}{p}\right) \\ \vdots \\ (-1)\sin\left(\frac{\pi.1}{p}\right) \\ 0 \end{bmatrix} \cos\left(\frac{\omega(T)}{T}t + \varphi_3\right) + C_5 \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix} \end{aligned} \tag{3.26}$$

if p is odd. The constants $C_1, \ldots, C_5$ *and* $\varphi_1, \ldots, \varphi_3$ *depend on the initial conditions.*

Remark 3.16. Under the assumption of the above proposition the consensus functional satisfies

$$V(\phi) = \frac{1}{p}\,[1 \cdots 1]\phi(0).$$

This follows from (2.8), taken into account that $V_0 = E_0$ if A is symmetric. Hence, under the conditions of the above proposition an *average consensus problem* is solved, in the sense that all components of a solution $x(\phi)(t)$ converge to the average of these components at the starting time, i.e., $\phi(0)$. Observe that $\phi(\theta)$, $\theta < 0$, has *no* influence on the limit reached.

Remark 3.17. Expression (3.23) reduces to the statement of Theorem 10 in [19] if $T \to 0+$.

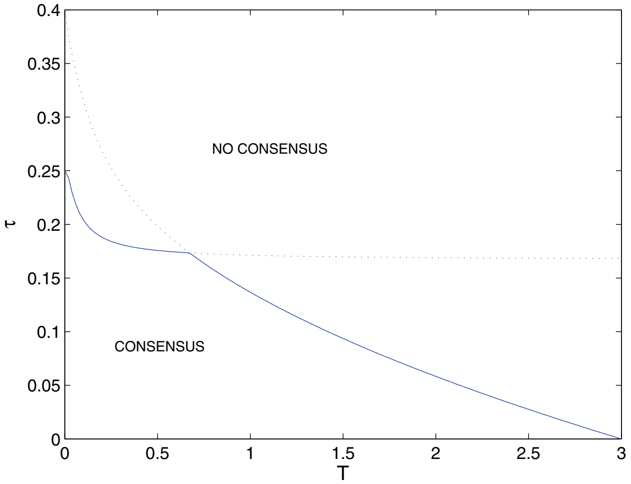


FIG. 1. *Boundary of the consensus regions of* (1.6)–(1.7) *with parameters* (4.1) *(solid curve). Boundaries of stability regions of* (4.2) *(dotted curves).*

Let us briefly compare the stationary solutions (3.21) with (3.25)–(3.26). In the former case, the directed "network topology" (a driver only reacts—with some delay—on its predecessor, and not the other way around) naturally leads to a backward *traveling* wave. In the latter case, one would from the symmetry of the coupling intuitively expect a *stationary* wave, where subsequent agents oscillate in antiphase. This is indeed the case for (3.25) which holds if p is even. However, if p is odd, such a solution is *incompatible* with the ring configuration, and (3.26) holds. If p is large, (3.26) can be seen as an approximation of a stationary wave with subsequent agents oscillating in antiphase that is compatible with the ring configuration.

4. Examples. As a first example we compute the consensus regions in the (T, τ) plane of system (1.6)–(1.7) with $n = 1$ and

$$A = \begin{bmatrix} -5 & 0 & 0 & 5 \\ 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 5 & -5 \end{bmatrix}. \tag{4.1}$$

The eigenvalues of this matrix are given by

$$\mu_1 = -6,\ \mu_2 = \bar{\mu}_3 = -3 + j,\ \mu_4 = 0.$$

An application of Theorem 3.9 yields the consensus region

$$T \in [0, 3),\ \tau \in [0, \tau^*(T)),$$

where the function $T \to \tau^*(T)$ is displayed in Figure 1 as a solid line. The dotted lines bound the "stability" regions of the auxiliary equations

$$s(1 + sT)e^{s\tau} - \mu_{1,2} = 0, \tag{4.2}$$

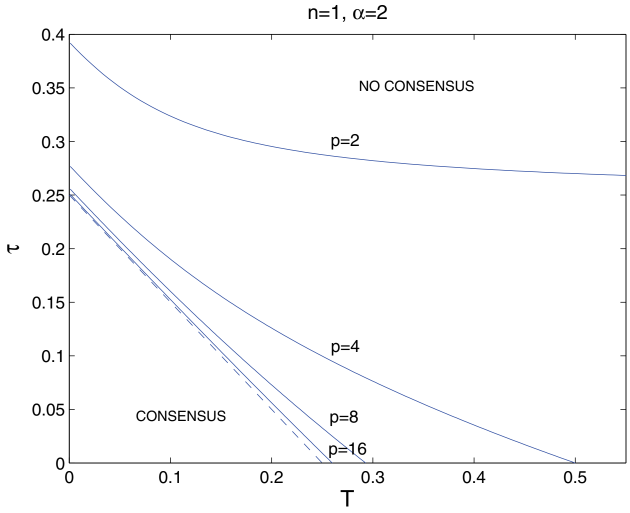


FIG. 2. *Boundary of the consensus region of a system satisfying* (3.18)*, with parameters* $\alpha = 2$ *and* $n = 1$*.*

which are described by Proposition 3.8. The stability region corresponding to μ_1 is unbounded as μ_1 is real and $n = 1$.

To illustrate the asymptotic behavior when the number of cars is large, we take a system satisfying condition (3.18) of Proposition 3.14. Figure 2 shows the consensus region in the (T, τ) plane for $\alpha = 2$, $n = 1$, and $p = 2^k$, $k = 1, \ldots, 4$. It follows from (3.19) that as $p \to \infty$, the boundary of the consensus region uniformly converges to the function

$$\tau_l^*(T) = \frac{1}{2\alpha} - nT,$$

indicated in Figure 2 with a dashed line.

Finally, we consider the system

$$\dot{v}_k(t) = \sum_{l=1}^{3} \alpha_{k,\lfloor k-l \rfloor} \int_0^\infty f(\theta)(v_{\lfloor k-l \rfloor}(t-\theta) - v_k(t-\theta))\, \mathrm{d}\theta, \quad k = 1, \ldots, 1000, \tag{4.3}$$

where f is given by (1.7), with $n = 2$. The parameters

$$\begin{aligned} &\alpha_{k,\lfloor k-1 \rfloor} \in [1, 5], \\ &\alpha_{k,\lfloor k-2 \rfloor} \in \left[0, \tfrac{3}{4}\alpha_{k,\lfloor k-1 \rfloor}\right], \\ &\alpha_{k,\lfloor k-3 \rfloor} \in \left[0, \tfrac{3}{4}\alpha_{k,\lfloor k-2 \rfloor}\right], \quad k = 1, \ldots, 1000, \end{aligned} \tag{4.4}$$

are randomly generated according to a *uniform distribution* over the above intervals. For 30 sets of parameters obtained in this way, the consensus region in the (T, τ) plane was computed. The results are displayed in Figure 3.

5. Other models. For general time-delay systems of retarded type with multiple constant delays and distributed delays with γ-distribution kernels, stability and/or consensus regions of equilibria in a two-parameter spaces can be computed semi-automatically by numerical continuation; see, for instance, [14] and the package DDE-BIFTOOL [7]. Such an approach involves the discretization of an infinite-dimensional

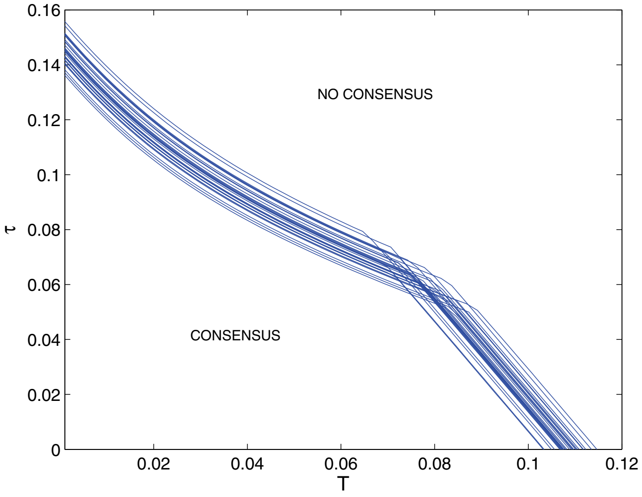


FIG. 3. *Consensus region of* (4.3)–(4.4) *for* 30 *different data sets.*

evolutionary operator, associated with the time-delay system, to compute the rightmost eigenvalues. Roughly speaking, the computation of the boundary of a stability or consensus region involves solving r eigenvalue problems of dimension $pq \times pq$, where p is the dimension of the system, q denotes the number of discretization points, and r is the number of points on the stability crossing curves where stability information is checked. When using Theorem 3.9 only one eigenvalue problem of dimension $p \times p$ needs to be solved to determine the complete stability region of (1.6)–(1.7) in the (T, τ) space. The underlying reason is that the structure of the system allowed a decomposition into small subproblems, which is apparent from the form of the characteristic equation (2.4).

Let us now take a brief look at the so-called optimal velocity models, also frequency encountered in the literature. The linearization around the equilibrium of the models studied in [1], respectively [5, 21, 22] and the references therein, takes the form

$$\tau_k \ddot{x}_k(t) + \dot{x}_k(t-\tau) = \alpha(x_{\lfloor k-1 \rfloor}(t-\tau) - x_k(t-\tau)), \qquad k = 1, \ldots, p, \tag{5.1}$$

respectively

$$\tau_k \ddot{x}_k(t) + \dot{x}_k(t) = \alpha(x_{\lfloor k-1 \rfloor}(t-\tau) - x_k(t-\tau)), \qquad k = 1, \ldots, p. \tag{5.2}$$

In both cases, x_k denotes the position of the kth vehicle. The left-hand side models the dynamics of the vehicle and the right-hand side is the reference velocity, which is a function of the distance to the preceding vehicle and models the behavior of the driver. Note that (5.1) and (5.2) can be generalized to

$$\begin{aligned} \ddot{x}_k(t) = \int_0^\infty f(\theta) \Bigg(\Bigg(\sum_{i=1}^{p-1} \alpha_{k,\lfloor k-i \rfloor} (x_{\lfloor k-i \rfloor}(t-\theta) - x_k(t-\theta)) \Bigg) \\ - \beta_k \dot{x}_k(t-\theta) \Bigg) \mathrm{d}\theta - \gamma_k \dot{x}_k(t), \qquad k = 1, \ldots, p, \end{aligned} \tag{5.3}$$

with f given by (1.7). General purpose tools for the stability and bifurcation analysis of time-delay systems like DDE-BIFTOOL can be applied directly to (5.3). However, if all vehicles have similar characteristics (but not necessarily the drivers), that is, $\beta_k \equiv \beta$, $\gamma_k \equiv \gamma$, then the characteristic equation can again be factorized:

$$\begin{aligned} 0 &= \det\left(s(s+\gamma)I - (A-\beta s I)\tfrac{e^{-s\tau}}{(1+sT)^n}\right) \\ &= \prod_{k=1}^{p}\left(s(s+\gamma) - \tfrac{(\mu_k-\beta s)e^{-s\tau}}{(1+sT)^n}\right), \end{aligned}$$

where A is given by (3.1) and $\mu_1, \ldots, \mu_p$ denote its eigenvalues. Also here, it is beneficial to exploit this decomposition into small subproblems, in particular if the number of vehicles is large.

6. Conclusions.

The stability analysis of a linear system including a γ-distributed delay with a gap for modeling traffic flow dynamics was considered. A complete characterization of the regions in the delay-parameter space, where a consensus is reached for all initial conditions, was obtained. In particular, by exploiting the structure of the system, analytical expressions were derived for the bounds on the parameters of the delay distribution. These expressions give rise to a fully automatic computation of the consensus region, whose complexity is determined by the computation of the eigenvalues of one matrix with dimensions equal to the number of vehicles. Some illustrative examples were presented.

From a theoretical point of view some stability theory for linear systems with γ-distributed delays was developed. As this type of distributed delays is characterized by kernels with an infinite support, which prohibits a full spectral decomposition of the solutions, the relation between the growth properties of the solutions and the roots of an appropriate characteristic equations was established via a comparison system with constant delays. Necessary and sufficient conditions for the realization of a consensus problem and an explicit construction of the consensus functional were provided, for both systems with constant and distributed delays.

Appendix A. Proof of Proposition 3.14.

The system matrix A, defined by (3.1), becomes circulant under condition (3.18). It is readily verified that its eigenvalues are given by

$$\mu_k = -\alpha + \alpha e^{j\frac{2\pi}{p}k}, \qquad k = 1, \ldots, p, \tag{A.1}$$

with corresponding eigenvectors

$$[(1+\mu_k/\alpha)^{p-1} \cdots (1+\mu_k/\alpha)^2 \ \ (1+\mu_k/\alpha) \ \ 1]^T. \tag{A.2}$$

The eigenvalues of A are real if and only if $p = 2$.

Let $q = p/2$ if p is even and $q = (p-1)/2$ otherwise. From (A.1) it follows that

$$\begin{aligned} |\mu_k| &= 2\alpha \sin\left(\tfrac{\pi}{p}k\right), \\ \angle(\mu_k) &= \tfrac{\pi}{2} + \tfrac{\pi}{p}k, \qquad k = 1, \ldots, q. \end{aligned} \tag{A.3}$$

Substituting these values in (3.14) yields

$$\tau^*(T) = \min_{k=1,\ldots,q} \frac{\left(\frac{\pi}{p}k - n\arctan(\omega_k(T))\right)T}{\omega_k(T)}, \tag{A.4}$$

where

$$|j\omega_k(T)(1+j\omega_k(T))^n| = 2\alpha T \sin\left(\frac{\pi}{p}k\right), \qquad k=1,\dots,q.$$

We have

$$\frac{k}{\omega_k(T)} = \frac{k\,|(1+j\omega_k(T))^n|}{2\alpha T\sin\left(\frac{\pi}{p}k\right)} \geq \frac{|(1+j\omega_1(T))^n|}{2\alpha T\sin\left(\frac{\pi}{p}\right)} = \frac{1}{\omega_1(T)}, \qquad k=1,\dots,q,$$

where we used $\omega_k(T) \geq \omega_1(T)$ and $\sin(kx) \leq k\sin(x)$ for $x \in [0,\pi/(2k)]$. This estimate and the fact that the function $x \to \arctan(x)/x$ is decreasing on $[0,\infty)$ lead to

$$\text{(A.5)} \qquad \begin{aligned} \frac{\left(\frac{\pi}{p}k - n\arctan(\omega_k(T))\right)T}{\omega_k(T)} &= \left(\frac{\pi k}{p\omega_k(T)} - \frac{n\arctan(\omega_k(T))}{\omega_k(T)}\right)T \\ &> \left(\frac{\pi}{p\omega_1(T)} - \frac{n\arctan(\omega_1(T))}{\omega_1(T)}\right)T \\ &= \frac{\left(\frac{\pi}{p} - n\arctan(\omega_1(T))\right)T}{\omega_1(T)}, \qquad k=1,\dots,q. \end{aligned}$$

From (A.4) and (A.5), one obtains (3.19).

Using (A.3) expression (3.15) becomes

$$T^* = \min_{k=1,\dots,q} \frac{\tan\left(\frac{\pi}{pn}k\right)}{2\alpha\sin\left(\frac{\pi}{p}k\right)\left(\cos\left(\frac{\pi}{pn}k\right)\right)^n}.$$

Taking into account that the function $x \mapsto \frac{\tan(x/n)}{\sin(x)\cos^n(x/n)}$ is increasing on $[0,\pi/2]$, one obtains (3.20).

Finally we consider the stationary solutions for $\tau = \tau^*(T)$. From the proof of Proposition 3.6, (A.1), and (A.3)–(A.5) it follows that the equation

$$s(1+sT)^n e^{s\tau^*(T)} \mp \mu_1 = 0$$

has solutions $s^\pm = \pm\frac{j\omega}{T}$, where $\omega = g_n\left(2\alpha T\sin\left(\frac{\pi}{p}\right)\right)$. From Lemma 3.4, (A.1), and Proposition 3.10 these solutions are eigenvalues of (2.7) with multiplicity one. As spelled out in the proof of Proposition 3.10 the corresponding eigenfunctions have the form $Z^\pm\, e^{\frac{\pm j\omega T}{T}}$, where $Z^\pm$ can be partitioned as $Z = [X^{\pm\ T}\ Y_0^{\pm\ T}\ \dots\ Y_{n-1}^{\pm\ T}]^T$, with

$$\text{(A.6)} \qquad e^{\frac{\pm j\omega t}{T}} X^\pm = e^{\frac{\pm j\omega t}{T}} \left[e^{\pm j\frac{2\pi(p-1)}{p}} \cdots e^{\pm j\frac{2\pi}{p}}\ 1\right]^T.$$

In the above we used (A.2). Note from Lemma 2.1 that (A.6) is a solution of (2.1). The solution (3.21) is a linear combination of (A.6), the solutions corresponding to the eigenvalues $\pm j\omega/T$, and $[1\cdots 1]^T$ (the eigenfunction corresponding to the zero eigenvalue). □

Appendix B. Proof of Proposition 3.15. The expressions (3.22) and (3.23) follow directly from Theorem 3.9, when taking into account that the eigenvalues of A are negative real.

Under condition (3.24) the eigenvalues of A are given by

$$\mu_k = -2\alpha + 2\alpha\cos\left(\frac{2\pi}{p}(k-1)\right), \qquad k=1,\dots,q,$$

where $q = (p+2)/2$ if p is even and $q = (p+1)/2$ if p is odd. All eigenvalues have multiplicity two, excepting μ_1 and, if p is even, $\mu_{\frac{p+2}{2}}$. The corresponding eigenvectors are

$$\begin{bmatrix} \cos\left(\frac{2\pi(k-1).(p-1)}{p}\right) \\ \vdots \\ \cos\left(\frac{2\pi(k-1).1}{p}\right) \\ 1 \end{bmatrix} \text{ and } \begin{bmatrix} \sin\left(\frac{2\pi(k-1).(p-1)}{p}\right) \\ \vdots \\ \sin\left(\frac{2\pi(k-1).1}{p}\right) \\ 0 \end{bmatrix}.$$

One can proceed as in the proof of Proposition 3.14.

Acknowledgments. The authors wish to thank the editors and the reviewers for their careful reading of the paper and their constructive comments.

REFERENCES

[1] M. Bando, K. Hasebe, K. Nakanishi, and A. Nakayama, *Analysis of optimal velocity model with explicit delay*, Phys. Rev. E, 58 (1998), pp. 5429–5435.

[2] R. E. Chandler, R. Herman, and E. W. Montroll, *Traffic dynamics: Analysis of stability in car following*, Oper. Res., 7 (1958), pp. 165–184.

[3] V. Chellaboina, W. M. Haddad, Q. Hui, and J. Ramakrishnan, *On systems state equipartitioning and semistability in network dynamical systems with arbitrary time-delays*, in Proceedings of the 45th IEEE Conference on Decision and Control, San Diego, CA, 2007, pp. 3461–3466.

[4] J. Chen, *Static output feedback stabilization for SISO systems and related problems: Solutions via generalized eigenvalues*, Control Theory Adv. Tech., 10 (1995), pp. 2233–2244.

[5] L. C. Davis, *Modifications of the optimal velocity traffic model to include delay due to driver reaction time*, Phys. A, 319 (2003), pp. 557–567.

[6] O. Diekmann, S. A. van Gils, S. M. Verduyn Lunel, and H.-O. Walther, *Delay Equations: Functional-, Complex-, and Nonlinear Analysis*, Appl. Math. Sci. 110, Springer-Verlag, New York, 1995.

[7] K. Engelborghs, T. Luzyanina, and G. Samaey, *DDE-BIFTOOL v.* 2.00*: A MATLAB Package for Bifurcation Analysis of Delay Differential Equations*, Technical report TW-330, Department of Computer Science, K. U. Leuven, Leuven, Belgium, 2001.

[8] M. Green, *How long does it take to stop?*, Transportation Human Factors, 2 (2000), pp. 195–216.

[9] K. Gu, V. L. Kharitonov, and J. Chen, *Stability of Time-Delay Systems*, Birkhäuser Boston, Boston, 2003.

[10] J. K. Hale and S. M. Verduyn Lunel, *Introduction to Functional Differential Equations*, Appl. Math. Sci. 99, Springer-Verlag, New York, 1993.

[11] J. Hale, *Theory of Functional Differential Equations*, Appl. Math. Sci. 3, Springer-Verlag, New York, 1977.

[12] D. Helbing, *Traffic and related self-driven many-particle systems*, Rev. Modern Phys., 73 (2001), pp. 1067–1141.

[13] A. Jadbabaie, J. Lin, and A. S. Morse, *Coordination of groups of mobile autonomous agents using nearest neighbor rules*, IEEE Trans. Automat. Control, 48 (2003), pp. 988–1001.

[14] T. Luzyanina and D. Roose, *Equations with distributed delays: Bifurcation analysis using computational tools for discrete delay equations*, Funct. Differ. Equ., 11 (2004), pp. 87–92.

[15] W. Michiels and S.-I. Niculescu, *Stability and Stabilization of Time-Delay Systems: An Eigenvalue Based Approach*, Adv. Des. Control 12, SIAM, Philadelphia, PA, 2007.

[16] C. I. Morărescu, *Qualitative Analysis of Distributed Delay Systems. Methodology and Algorithms*, Ph.D. thesis, University of Bucharest, Romania/Université de Technologie de Compiègne, France, 2006.

[17] C. I. Morărescu, S.-I. Niculescu, and W. Michiels, *Asymptotic stability of some distributed delay systems: An algebraic approach*, International Journal of Tomography and Statistics, 7 (2007), pp. 128–132.

[18] S.-I. Niculescu and C. T. Abdallah, *Delay effects on static output feedback stabilization*, in Proceedings of the 39th IEEE Conference on Decision and Control, Sydney, Australia, 2000, pp. 2811–2816.

[19] R. Olfati-Saber and R. M. Murray, *Consensus problems in networks of agents with switching topology and time-delays*, IEEE Trans. Automat. Control, 49 (2004), pp. 1520–1533.

[20] R. Olfati-Saber, A. Fax, and R. M. Murray, *Consensus and cooperation in networked multi-agent systems*, Proc. IEEE, 95 (2007), pp. 215–233.

[21] G. Orosz, B. Krauskopf, and R. E. Wilson, *Bifurcations and multiple traffic jams in a car-following model with reaction-time delay*, Phys. D, 211 (2005), pp. 277–293.

[22] G. Orosz and G. Stépán, *Subcritical Hopf bifurcations in a car-following model with reaction-time delay*, Proc. R. Soc. London Ser. A Math. Phys. Eng. Sci., 462 (2006), pp. 2643–2670.

[23] W. Ren, R. W. Beard, and E. Atkins, *Information consensus in multivehicle cooperative control*, IEEE Control Systems Magazine, 27 (2007), pp. 71–82.

[24] R. W. Rothery, *Transportation Research Board (TRB) Special Report* 165, in Traffic Flow Theory, 2nd ed., N. H. Gartner, C. J. Messner, and A. J. Rathi, eds., Transportation Research Board, Washington, DC, 1998.

[25] R. Sipahi and S.-I. Niculescu, *Analytical stability study of a deterministic car following model under multiple delay interactions*, in Proceedings of the 6th IFAC Workshop on Time Delay Systems, L'Aquila, Italy, 2006.

[26] R. Sipahi and S.-I. Niculescu, *Some remarks on the characterization of delay interactions in deterministic car following models*, in Proceedings of the 17th Annual Symposium MTNS, Kyoto, Japan, 2006.

[27] R. Sipahi, F. M. Atay, and S.-I. Niculescu, *Stability of traffic flow behavior with distributed delays modeling the memory effects of the drivers*, SIAM J. Appl. Math., 68 (2007), pp. 738–759.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 1, pp. 102–133

© 2009 Society for Industrial and Applied Mathematics

AVERAGE CONSENSUS WITH PACKET DROP COMMUNICATION*

FABIO FAGNANI† AND SANDRO ZAMPIERI‡

Abstract. Average consensus consists in the problem of determining the average of some quantities by means of a distributed algorithm. It is a simple instance of problems arising when designing estimation algorithms operating on data produced by sensor networks. Simple solutions based on linear estimation algorithms have already been proposed in the literature and their performance has been analyzed in detail. If the communication links which allow the data exchange between the sensors have some loss, then the estimation performance will degrade. In this contribution the performance degradation due to this data loss is evaluated.

Key words. average consensus, Cayley graphs, mean square analysis, packet drop, rate of convergence

AMS subject classifications. 05C80, 68W15, 68W20, 93A15

DOI. 10.1137/060676866

1. Introduction.

Average consensus problems have been widely studied in recent years [20, 12, 18, 2, 15, 9, 16], both in the context of coordination of mobile autonomous vehicles and in the context of distributed estimation. In fact, average consensus can be considered a simple paradigm for designing estimation algorithms implemented on sensor networks and working in a distributed way. More precisely, we assume in this setup that all sensors independently measure the same quantity with some error due to noise. A simple way to improve the estimate is to average all the measures. To do this, the sensors need to exchange their information. Energy limitations force transmission to take place directly along nearby sensors and also impose bounds on the amount of data an agent can process. A global description of the allowed exchange of information can be given by a directed graph in which the sensors are the nodes and in which an edge from agent i to agent j represents the possibility for i to send information to j. Algorithms which allow us to obtain this average are called average consensus algorithms. The performance of an average consensus algorithm may be measured by the speed of convergence toward the average. In [15] a simple algorithm is proposed which is based on a linear dynamical system. Moreover, in [15, 3] the relation between the performance of this algorithm and the degree of connectivity of the graph is also evaluated. In [20, 11, 13, 18] variations of this algorithm which handle time-varying communication graphs are considered. Since in these cases the analysis proposed is essentially a worst case analysis, the performance evaluation can be rather conservative. Different results can be obtained if the graphs vary in time randomly [2, 3]. In fact, randomly time-varying graphs typically yield improved performance.

In this paper we consider a more realistic model of the data exchange. In fact, in many practical applications, the data exchange between sensors takes place over a wireless communication network leading to the possibility that some packets get

*Received by the editors December 6, 2006; accepted for publication (in revised form) July 30, 2008; published electronically February 11, 2009.
http://www.siam.org/journals/sicon/48-1/67686.html

†Dipartimento di Matematica, Politecnico di Torino, C.so Duca degli Abruzzi, 24, 10129 Torino, Italy (fabio.fagnani@polito.it).

‡Department of Information Engineering, Università di Padova, Via Gradenigo 6/a, 35131 Padova, Italy (zampi@dei.unipd.it).

lost during the transmission. In this contribution this phenomenon is modeled by assuming that at every time instant the transmission of a number from one sensor to another can occur with a certain probability, and so there is a certain probability that the link will fail and the data will be lost. We can expect that this will produce a performance degradation. The main objective of this contribution is to provide some instruments that allow us to quantify this degradation as a function of the probability of the link failure. The problem is similar to the one considered in [9] where a more limited class of random graphs were considered and where only the convergence of the algorithm was considered. We have recently realized that other researchers [17] have independently studied similar problems; their results, however, are different from ours.

In section 2, after recalling classical average consensus algorithms, we propose two different adaptations of such algorithms which can cope with lossy links: the biased and the balanced compensation methods. The essential difference between the two methods is that in the biased version, local averaging weights at each node are kept fixed while, in the balanced case, weights are scaled depending on the available data at every instant. Both algorithms will be shown to converge (almost surely and in mean square sense) to a consensus value which in general may not coincide with the average of the initial states. For both cases, performance degradation will be analyzed through two figures showing the rate of convergence and the asymptotic displacement from the average consensus. Analysis will always be carried out in a mean square sense. Analysis of the degradation of the convergence rate is undertaken in section 3, where the problem is reduced to finding the largest eigenvalue of a suitable linear operator $\mathcal{L}$ acting on a space of N^2 dimensions (where N is the number of agents). This reduction, besides giving an important theoretical characterization, is amenable to efficient numerical analysis simulations. Sections 4 and 5 are devoted to the case when the network possesses symmetries, in particular when it can be modeled by an Abelian Cayley graph. In this case, the operator $\mathcal{L}$ can actually be substituted with an N-dimensional operator. This allows us to obtain deeper analytical results and, in particular, to obtain explicit solutions in special important cases (e.g., complete graph, cycle graph, and hypercube graph). A comparison of the two methods shows that, at least in some examples, the balanced method presents a better rate of convergence. Finally, in section 6, we analyze the asymptotic displacement from the average consensus due to packet drop and we prove that for the Abelian Cayley case, this displacement is infinitesimal in the number of agents for both methods. We will also show that with respect to the asymptotic displacement the biased method outperforms the other.

2. Problem formulation. We assume that we have N agents. Each agent i measures a quantity $d_i \in \mathbb{R}$ and at each time instant t it can transmit a real number to some agents. The data exchange can be described by a directed graph $\mathcal{G}$ with vertices $\{1, \dots, N\}$, in which there is an edge (j, i) if and only if the agent j can send data to agent i. The objective is to find a distributed algorithm which allows the agents to obtain a shared estimate of the average of the d_i's. An efficient algorithm solving this problem consists in the dynamic system

$$x_i(t+1) = \sum_{j=1}^{N} P_{ij} x_j(t), \qquad x_i(0) = d_i,$$

where P is a suitable matrix such that $P_{ij} = 0$ if (j, i) is not an edge in $\mathcal{G}$. We assume that $\mathcal{G}$ always includes all the self loop edges (i, i), meaning that each agent i has access to its own data. More compactly we can write

$$x^+ = Px, \qquad x(0) = d, \tag{1}$$

where $x, d \in \mathbb{R}^N$ and where x^+ is a shorthand notation for $x(t+1)$. According to this algorithm, the agent i needs to receive the value of $x_j(t)$ from the agent j to update the value of $x_i(t)$ only if $P_{ij} \neq 0$. In this case, we say that the agents reach the *consensus*, if for any initial condition $x(0) \in \mathbb{R}^N$, the closed loop system (1) yields

$$\lim_{t\to\infty} x(t) = \mathbf{1}\alpha, \tag{2}$$

where $\mathbf{1} := (1, \dots, 1)^*$ and where α is a scalar depending on $x(0)$ and P. Moreover, if α coincides with the average $N^{-1} \sum_{i=1}^N d_i = N^{-1}\mathbf{1}^* x(0)$, then we say that the agents reach the *average consensus*.

To make the concepts more precise it is useful to recall some notation and results on directed graphs (the reader can further refer to textbooks on graph theory such as [8] or [5]). Fix a directed graph $\mathcal{G}$ with a set of vertices V and a set of arcs $\mathcal{E} \subseteq V \times V$. The adjacency matrix E is a $\{0, 1\}$-valued square matrix indexed by the elements in V defined by letting $E_{ij} = 1$ if and only $(i, j) \in \mathcal{E}$. Define the out-degree of a vertex j as $\operatorname{outdeg}(j) := \sum_i E_{ij}$ and the in-degree of a vertex i as $\operatorname{indeg}(i) := \sum_j E_{ij}$. A graph is called in-regular (resp., out-regular) of degree k if each vertex has in-degree (resp., out-degree) equal to k. A path in $\mathcal{G}$ consists of a sequence of vertices $i_1 i_2 \dots i_r$ such that $(i_\ell, i_{\ell+1}) \in \mathcal{E}$ for every $\ell = 1, \dots, r-1$; i_1 (resp., i_r) is said to be the initial (resp., terminal) vertex of the path. A path is said to be closed if the initial and the terminal vertices coincide. A vertex i is said to be connected to a vertex j if there exists a path with initial vertex i and terminal vertex j. A directed graph is said to be connected if, given any pair of vertices i and j, either i is connected to j or j is connected to i. A directed graph is said to be strongly connected if, given any pair of vertices i and j, i is connected to j.

With an $N \times N$ matrix P we associate a directed graph $\mathcal{G}_P$ with a set of vertices $\{1, \dots, N\}$ in which there is an arc from j to i whenever the element $P_{ij} \neq 0$. The graph $\mathcal{G}_P$ is said to be the *communication graph* associated with P. Conversely, given any directed graph $\mathcal{G}$ with the set of vertices $\{1, \dots, N\}$, we say that a matrix P is *compatible* with $\mathcal{G}$ if $\mathcal{G}_P$ is a subgraph of $\mathcal{G}$. After introducing this notation we can make the consensus problem more precise. We say that the (average) consensus problem is solvable on a graph $\mathcal{G}$ if there exists a matrix P compatible with $\mathcal{G}$ and solving the (average) consensus problem.

As shown in [18, 15, 3], if $\mathcal{G}$ is strongly connected, it is always possible to choose P so as to obtain the consensus. Indeed, if P is a stochastic matrix (namely, $P_{ij} \geq 0$ for every i, j and $P\mathbf{1} = \mathbf{1}$), $\mathcal{G}_P$ is strongly connected, and $P_{ii} > 0$ for some i, then P solves the consensus problem. To obtain average consensus P needs to satisfy an extra condition: It must be doubly stochastic ($\mathbf{1}^* P = \mathbf{1}^*$). If $\mathcal{G}$ is strongly connected, a P also satisfying this last condition can be found even if the construction becomes, in general, more involved. There is an important case when the construction of such a P is quite simple—when all agents have the same out- and in-degree ν (without considering self loops). In this case, we can simply choose $P = kI + (1-k)\nu^{-1}E$ for any $k \in]0, 1[$. Undirected graphs are clearly an example which fits into this case and P in this case is actually symmetric.

In the following we will give an elementary example which casts the average consensus problem into the topic of distributed estimation.

Example 1 (estimation from distributed measures [14, 2]). Assume we have N sensors which measure a quantity $z \in \mathbb{R}$. However, due to noise, each sensor obtains different measures $y_i = z + v_i$, where v_i are independent random variables with zero mean and the same variance. It is well known that the average

$$\alpha = N^{-1} \sum_{i=1}^{N} y_i$$

provides the best possible linear estimate of z (in the sense of the minimum mean square error) from y_i. Running an average consensus problem with initial conditions $x_i(0) = y_i$ will lead to a distributed computation of α by every agent.

2.1. Packet drop consensus algorithms. We start from a fixed graph $\mathcal{G}$ and we assume that on each edge (j, i) of $\mathcal{G}$, communication from the node j to the node i can occur with some probability p. In order to describe this model more precisely, we introduce the family of independent binary random variables $L_{ij}(t)$, $t \in \mathbb{N}$, $i, j = 1, \dots, N$, $i \neq j$, such that

$$\mathbb{P}[L_{ij}(t) = 1] = p, \qquad \mathbb{P}[L_{ij}(t) = 0] = 1 - p.$$

We emphasize the fact that independence is assumed among all $L_{ij}(t)$ as i, j and t vary. Let E be the adjacency matrix of $\mathcal{G}$, and let $H := E - I$. Consider the random matrix $\bar{E}(t) = I + \bar{H}(t)$, where $\bar{H}_{ij}(t) = H_{ij} L_{ij}(t)$. Clearly, $\bar{E}(t)$ is the adjacency matrix of a random graph $\bar{\mathcal{G}}(t)$ obtained from $\mathcal{G}$ by deleting the edge (i, j) when $L_{ij}(t) = 0$.

In this paper we will propose consensus strategies compatible with the random varying communication graphs $\bar{\mathcal{G}}(t)$; they will consist of a sequence of random stochastic matrices $P(t)$ such that $\mathcal{G}_{P(t)} \subseteq \bar{\mathcal{G}}(t)$ for all t.

Our construction always starts from the choice of a stochastic matrix P adapted to $\mathcal{G}$ yielding average consensus and we modify it in a way to compensate for the lack of some data. There is, in principle, more than one way to obtain this. We will propose two solutions. In the first, which will be called the *biased compensation method*, each agent, in updating the estimate of the average, adds the weights of the unavailable data to the weight it assigns to its own old estimate. In the second, which will be called the *balanced compensation method*, the compensation for the lack of data is done by modifying all the weights in a more balanced way. We want to emphasize the fact that we are assuming all agents to be time synchronized. As a consequence, at every time instant t, any agent i knows which data he has received; this means that agent i knows the value of $L_{ij}(t)$ for every neighbor j.

The biased compensation method. We consider the following updating law:

$$x_i(t+1) = \left(P_{ii} + \sum_{j \neq i} (1 - L_{ij}(t)) P_{ij} \right) x_i(t) + \sum_{j \neq i} L_{ij}(t) P_{ij} x_j(t).$$

According to this strategy, the agent i, in computing the new estimate, compensates for the loss of data by accumulating the weights of the lost data with the weight assigned to its previous estimate. Intuitively, according to this method, the agent i

substitutes the unavailable $x_j(t)$ with $x_i(t)$ in the consensus algorithm. If we define the random matrices $D(t), Q(t)$ as

$$D_{ij}(t) := \begin{cases} P_{ii} + \sum_{j \neq i}(1 - L_{ij}(t))P_{ij} = 1 - \sum_{h \neq i} L_{ih}(t)P_{ih} & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases}$$

and

$$Q_{ij}(t) := \begin{cases} 0 & \text{if } i = j, \\ L_{ij}(t)P_{ij} & \text{if } i \neq j, \end{cases}$$

then we can describe this method through the stochastic system

$$x(t+1) = P(t)x(t), \tag{3}$$

where

$$P(t) := D(t) + Q(t).$$

The balanced compensation method. As opposed to the previous method, here we prefer to distribute the weights equally between the available data. The updating equation is thus

$$x_i(t+1) = \frac{1}{P_{ii} + \sum_{j \neq i} L_{ij}(t)P_{ij}} \left(P_{ii}x_i(t) + \sum_{j \neq i} L_{ij}(t)P_{ij}x_j(t) \right).$$

In this case it is convenient to define, for $i = 1, \ldots, N$, the binary random variable L_{ii} which is equal to 1 with probability 1. In this way, by defining

$$\nu_i(t) = \sum_{j=1}^{N} L_{ij}(t)P_{ij}$$

and introducing the diagonal matrix $D(t)$ having

$$D_{ii}(t) := \frac{1}{\nu_i(t)}$$

and the matrix $Q(t)$ such that

$$Q_{ij}(t) := L_{ij}(t)P_{ij},$$

we can more compactly write

$$x(t+1) = P(t)x(t)$$

with

$$P(t) := D(t)Q(t).$$

Our goal will be to evaluate the asymptotic behavior of system (3) in the two cases. The following result shows that, for both of the above methods, communication failures will never prevent us from reaching consensus. The proof is a simple consequence of Theorem 6 in [4] and is a particular instance of Corollary 3.2 in [6].

THEOREM 2.1. *Assume that P achieves the consensus and that $P_{ii} > 0$ for every i. Then, both the biased compensation method and the balanced compensation method yield consensus almost surely; namely,* (2) *almost surely holds for any initial condition $x(0)$ with an α which is general a random variable. Moreover, the convergence in* (2) *also holds in the mean square sense; namely,*

$$\lim_{t\to\infty} \mathbb{E}\left[||x(t) - \mathbf{1}\alpha||^2\right] = 0, \tag{4}$$

where $\mathbb{E}[\cdot]$ is the expected value and $||\cdot||$ is the 2-norm in $\mathbb{R}^N$.

Notice that the random variable α depends linearly on the initial condition $x(0)$. In other terms there exists an N-dimensional random vector v such that $\alpha = v^* x(0)$.

In spite of the previous theorem, we do expect a performance degradation due to communication failures. Degradation will show up in two ways; first, as a diminished convergence speed of the limit (4) and, second, as the deviation of the random variable α from the average. The aim of this paper is to quantify such a degradation. The next section will focus on the speed of convergence.

3. Mean square analysis. In this section we assume we have fixed a matrix P adapted to the graph $\mathcal{G}$ yielding consensus and such that $P_{ii} > 0$ for every node i. We then undertake a mean squared analysis of our stochastic models and we characterize their asymptotic rate of convergence. Precisely, our aim is to evaluate the exponential rate of convergence to 0 of $\mathbb{E}[||x(t) - \mathbf{1}\alpha||^2]$. We start with a preliminary result. Let

$$x_A(t) := \frac{1}{N}\sum_{i=1}^{N} x_i(t) = \frac{1}{N}\mathbf{1}^* x(t), \tag{5}$$

which coincides with the average of the current states.

PROPOSITION 3.1. *Assume that we have almost sure consensus, namely, that $x(t) \to \alpha\mathbf{1}$ almost surely. Then,*

$$\mathbb{E}[||x(t) - \mathbf{1}x_A(t)||^2] \leq \mathbb{E}[||x(t) - \mathbf{1}\alpha||^2] \leq (1 + \sqrt{N})^2 \mathbb{E}[||x(t) - \mathbf{1}x_A(t)||^2]. \tag{6}$$

Proof. From

$$x(t) - \mathbf{1}x_A(t) = (I - N^{-1}\mathbf{1}\mathbf{1}^*)x(t) = (I - N^{-1}\mathbf{1}\mathbf{1}^*)(x(t) - \mathbf{1}\alpha),$$

we obtain

$$||x(t) - \mathbf{1}x_A(t)|| \leq ||x(t) - \mathbf{1}\alpha||\,.$$

This proves the left inequality.

The following identity holds for every t and s:

$$x(t) - x(t+s) = (I - P(t+s-1)\cdots P(t))\,(I - N^{-1}\mathbf{1}\mathbf{1}^*)x(t)\,.$$

Using the fact that for any stochastic matrix P, $||P|| \leq \sqrt{N}$, we obtain that

$$||x(t) - x(t+s)|| \leq (1 + \sqrt{N})||x(t) - \mathbf{1}x_A(t)||\,. \tag{7}$$

Letting $s \to \infty$ and taking the average of this yields the right inequality. □

The proposition shows that $\mathbb{E}[||x(t) - \mathbf{1}\alpha||^2]$ and $\mathbb{E}[||x(t) - \mathbf{1}x_A(t)||^2]$ have the same exponential rate of convergence to zero or, in other words, that, for any initial condition $x(0)$, we have that

$$\limsup_{t\to+\infty} \mathbb{E}[||x(t) - \mathbf{1}\alpha||^2]^{1/t} = \limsup_{t\to+\infty} \mathbb{E}[||x(t) - \mathbf{1}x_A(t)||^2]^{1/t}\,.$$

For this reason, in what follows we will study the right-hand expression, which turns out to be simpler to analyze. In order to have a single figure not dependent on the initial condition, we will concentrate on this worst case exponential rate of convergence:

$$R := \sup_{x(0)} \limsup_{t\to+\infty} \mathbb{E}[||x(t) - \mathbf{1}x_A(t)||^2]^{1/t}\,.$$

Remark 1. Some of the considerations carried out above hold true even without a priori knowledge of almost sure consensus. This is true for (7) in the proof of Proposition 3.1 which, in any case, yields

$$(\mathbb{E}||x(t) - x(t+s)||^2)^{1/2} \leq (1+\sqrt{N})(\mathbb{E}||x(t) - \mathbf{1}x_A(t)||^2)^{1/2}\,. \tag{8}$$

Hence, from the simple knowledge that $\mathbb{E}||x(t) - \mathbf{1}x_A(t)||^2$ converges to 0, we can deduce that $x(t)$ is a Cauchy sequence and so, for completeness arguments, $x(t)$ converges in mean square to some random vector $x(\infty)$. Notice, moreover, that for any vector $\zeta \in \mathbb{R}^N$ orthogonal to $\mathbf{1}$ we have that

$$|\zeta^* x(t)| = |\zeta^*(x(t) - \mathbf{1}x_A(t))| \leq ||\zeta||||x(t) - \mathbf{1}x_A(t)|| \longrightarrow 0.$$

Since $\zeta^* x(t) \longrightarrow \zeta^* x(\infty)$, for the limit uniqueness we have that $\zeta^* x(\infty) = 0$. This implies that $x(\infty) = \mathbf{1}\alpha$ for some random variable α. Hence, convergence of $\mathbb{E}||x(t) - \mathbf{1}x_A(t)||^2$ yields consensus in the mean square sense.

In order to study the behavior of $\mathbb{E}[||x(t) - \mathbf{1}x_A(t)||^2]$, the following characterization turns out to be very useful. Indeed, notice that

$$\mathbb{E}[||x(t) - \mathbf{1}x_A(t)||^2] = \mathbb{E}[x^*(t)(I - N^{-1}\mathbf{1}\mathbf{1}^*)x(t)] = x^*(0)\Delta(t)x(0),$$

where

$$\Delta(t) := \mathbb{E}[P(0)^*P(1)^* \cdots P(t-1)^*(I - N^{-1}\mathbf{1}\mathbf{1}^*)P(t-1)\cdots P(1)P(0)],$$

if $t \geq 1$ and where $\Delta(0) := I - N^{-1}\mathbf{1}\mathbf{1}^*$. Therefore we have that

$$R = \max_{ij} \limsup_{t\to+\infty} \Delta(t)_{ij}^{1/t}\,.$$

We now study the evolution of the matrices $\Delta(t)$.

Notice first that

$$\begin{aligned}
&\Delta(t+1)\\
&= \mathbb{E}[P(0)^*P(1)^* \cdots P(t-1)^*P(t)^*(I - N^{-1}\mathbf{1}\mathbf{1}^*)P(t)P(t-1)\cdots P(1)P(0)]\\
&= \mathbb{E}[\mathbb{E}[P(0)^*P(1)^* \cdots P(t-1)^*P(t)^*(I - N^{-1}\mathbf{1}\mathbf{1}^*)P(t)P(t-1)\cdots P(1)P(0)|P(0)]]\\
&= \mathbb{E}[P(0)^*\mathbb{E}[P(1)^* \cdots P(t-1)^*P(t)^*(I - N^{-1}\mathbf{1}\mathbf{1}^*)P(t)P(t-1)\cdots P(1)]P(0)]\\
&= \mathbb{E}[P(0)^*\Delta(t)P(0)],
\end{aligned}$$

where the last equality follows from the fact that, since the random matrices $P(t)$ are independent and identically distributed, the two sequences of random matrices $(P(0),\dots,P(t-1))$ and $(P(1),\dots,P(t))$ have the same probability distribution.

It is convenient to introduce the linear operator $\mathcal{L}:\mathbb{R}^{N\times N}\to\mathbb{R}^{N\times N}$ defined by

$$\mathcal{L}(\Delta)=\mathbb{E}[P(0)^*\Delta P(0)].$$

In this way $\Delta(t)$ is governed by the recursive relation

$$\Delta(t+1)=\mathcal{L}(\Delta(t)).$$

If we consider now the reachable subspace $\mathcal{R}$ of the pair $(\mathcal{L},\Delta(0))$, namely, the smallest $\mathcal{L}$-invariant subspace of $\mathbb{R}^{N\times N}$ containing $\Delta(0)$, we clearly have that

$$R=\max\{|\lambda|\ :\ \lambda \text{ eigenvalue of } \mathcal{L}_{|\mathcal{R}}\},$$

where $\mathcal{L}_{|\mathcal{R}}$ denotes the restriction of the operator $\mathcal{L}$ to the invariant subspace $\mathcal{R}$.

The previous proposition implies that, under mild hypotheses, $\mathbf{L}^*$ is an irreducible aperiodic stochastic matrix row and therefore the eigenvalue 1 has algebraic multiplicity 1.

The operator $\mathcal{L}$ has many interesting properties which have been studied in [6] in a more general context. It has been shown in particular that $\mathcal{L}$ can be interpreted as an aperiodic row-stochastic operator. As a consequence, 1 is an eigenvalue of algebraic multiplicity one. It is easy to find a corresponding eigenvector. Notice indeed that $x(0)^*\mathcal{L}^t(\Delta)x(0)=\mathbb{E}[x(t)^*\Delta x(t)]$. Since $x(t)\to\mathbf{1}v^*x(0)$ in mean square sense, it follows that

$$\mathbb{E}[x(t)^*\Delta x(t)]\to x(0)^*\mathbf{1}^*\Delta\mathbf{1}\mathbb{E}[vv^*]x(0).$$

As a consequence,

$$\lim_{t\to+\infty}\mathcal{L}^t(\Delta)=(\mathbf{1}^*\Delta\mathbf{1})\mathbb{E}[vv^*].$$

In particular, $\mathcal{L}(\mathbb{E}[vv^*])=\mathbb{E}[vv^*]$. Clearly the reachability subspace $\mathcal{R}$ will be contained in the subspace generated by the eigenvectors different from $\mathbb{E}[vv^*]$.

In what follows we will write the operator $\mathcal{L}$ in a more explicit form. This will allow us to determine R numerically. To do this we now need to study the two cases separately.

3.1. The biased compensation method. For any matrix M we will denote diag (M) as the diagonal matrix with the same diagonal elements of M and out $(M):=M-\text{diag }(M)$ which is out-diagonal, namely, has zero diagonal elements.

PROPOSITION 3.2. *The sequence of matrices $\Delta(t)$ satisfies the recursive relation*

$$\begin{aligned}\Delta^+&=[(1-p)I+pP]^*\Delta[(1-p)I+pP]\\&\quad+p(1-p)\text{diag }\{\text{out }(P)\text{out }(P)^*\text{diag }(\Delta)+\text{out }(P)^*\text{diag }(\Delta)\text{out }(P)\}\\&\quad-p(1-p)\big\{\text{diag }(\Delta)\text{out }(\tilde{P})+\text{out }(\tilde{P})^*\text{diag }(\Delta)\big\},\end{aligned}\tag{9}$$

where the matrix $\tilde{P}$ is defined by letting $\tilde{P}_{ij}:=P_{ij}^2$.

Proof. Let $D:=D(0)$ and $Q:=Q(0)$. Notice, preliminarily, that $\mathbb{E}[Q]=p\,\text{out }(P)$ and that $\mathbb{E}[D]=(1-p)I+p\text{diag }(P)$. Notice, moreover, that

$$\mathbb{E}[D_{ii}D_{jj}]=\begin{cases}(1-p+pP_{ii})(1-p+pP_{jj}) & \text{if } i\neq j,\\ (1-p+pP_{ii})^2+p(1-p)\sum_{k\neq i}P_{ik}^2 & \text{if } i=j,\end{cases}\tag{10}$$

and that

$$\mathbb{E}[D_{ii}Q_{ij}] = (1-p+pP_{ii})pP_{ij} - p(1-p)P_{ij}^2. \tag{11}$$

Notice now that

$$\Delta^+ = \mathbb{E}[D\Delta D] + \mathbb{E}[D\Delta Q] + \mathbb{E}[Q^*\Delta D] + \mathbb{E}[Q^*\Delta Q].$$

Using (10) we obtain that

$$\begin{aligned}&\mathbb{E}[D\Delta D]_{ij}\\ &= \mathbb{E}\left[D_{ii}D_{jj}\right]\Delta_{ij} = \begin{cases} (1-p+pP_{ii})(1-p+pP_{jj})\Delta_{ij} & \text{if } i \neq j,\\ (1-p+pP_{ii})^2\Delta_{ii} + p(1-p)\left(\sum_{k\neq i} P_{ik}^2\right)\Delta_{ii} & \text{if } i = j.\end{cases}\end{aligned}$$

More compactly, we can write

$$\begin{aligned}\mathbb{E}[D\Delta D] = &\ [(1-p)I + p\text{diag }(P)]\Delta[(1-p)I + p\text{diag }(P)]\\ &+ p(1-p)\text{diag}\{\text{out}(P)\text{out }(P)^*\}\text{diag }(\Delta).\end{aligned}$$

Notice now that

$$\mathbb{E}[D\Delta Q]_{ii} = \sum_{k\neq i}\mathbb{E}\left[D_{ii}\Delta_{ik}Q_{ki}\right] = \mathbb{E}\left[D_{ii}\right]\sum_{k\neq i}\Delta_{ik}\mathbb{E}\left[Q_{ki}\right] = p(1-p+pP_{ii})\sum_{k\neq i}\Delta_{ik}P_{ki}.$$

If instead $i \neq j$, then, using (11), we obtain

$$\begin{aligned}\mathbb{E}[D\Delta Q]_{ij} &= \sum_{k\neq j}\mathbb{E}\left[D_{ii}\Delta_{ik}Q_{kj}\right] = \mathbb{E}\left[D_{ii}\right]\sum_{\substack{k\neq i\\ k\neq j}}\Delta_{ik}\mathbb{E}\left[Q_{kj}\right] + \mathbb{E}\left[D_{ii}\Delta_{ii}Q_{ij}\right]\\ &= p(1-p+pP_{ii})\sum_{\substack{k\neq i\\ k\neq j}}\Delta_{ik}P_{kj} + p(1-p+pP_{ii})P_{ij}\Delta_{ii} - p(1-p)P_{ij}^2\Delta_{ii}.\end{aligned}$$

More compactly, we can write

$$\mathbb{E}[D\Delta Q] = p[(1-p)I + p\text{diag }(P)]\Delta\text{out }(P) - p(1-p)\text{diag }(\Delta)\text{out }(\tilde{P}).$$

Finally, observe that

$$\begin{aligned}\mathbb{E}[Q^*\Delta Q]_{ii} &= \sum_{\substack{h\neq i\\ k\neq i}}\mathbb{E}\left[Q_{hi}\Delta_{hk}Q_{ki}\right] = p^2\sum_{\substack{h\neq i\ k\neq i\\ h\neq k}}P_{hi}\Delta_{hk}P_{ki} + p\sum_{h\neq i}P_{hi}^2\Delta_{hh}\\ &= p^2\sum_{\substack{h\neq i\\ k\neq i}}P_{hi}\Delta_{hk}P_{ki} + p(1-p)\sum_{h\neq i}P_{hi}^2\Delta_{hh}.\end{aligned}$$

If instead $i \neq j$, then

$$\mathbb{E}[Q^*\Delta Q]_{ij} = \sum_{\substack{h\neq i\\ k\neq j}}\mathbb{E}\left[Q_{hi}\Delta_{hk}Q_{kj}\right] = p^2\sum_{\substack{h\neq i\\ k\neq j}}P_{hi}\Delta_{hk}P_{kj}.$$

More compactly, we can write

$$E[Q^*\Delta Q] = p^2\text{out }(P)^*\Delta\text{out }(P) + p(1-p)\text{diag }\{\text{out }(P)^*\text{diag }(\Delta)\text{out }(P)\}.$$

Putting all the pieces together we obtain relation (9). □

Following previous considerations, we are interested in evaluating the eigenvalues of the linear map which furnishes $\Delta(t+1)$ from $\Delta(t)$. These matrices are symmetric and so the linear dynamic system described in the previous proposition has a state space of dimension $\frac{N(N-1)}{2}$.

Remark 2. Numerical algorithms can clearly be employed to evaluate such eigenvalues. The following is a concrete way to achieve this. Given a matrix $A \in \mathbb{R}^{N\times N}$, we define vect$(A)$ to be the N^2 column vector having $A_{i,j}$ in position $(i-1)N+j$. Moreover, let

$$M := \begin{bmatrix} e_1e_1^* & 0 & \cdots & 0 \\ 0 & e_2e_2^* & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e_Ne_N^* \end{bmatrix}, \tag{12}$$

where e_i is the vector with all zeros except for a 1 in the ith position. This matrix is such that $\text{vect}(\text{diag }(A)) = M\text{vect}(A)$. Finally, notice that $\text{vect}(ABC) = (C^* \otimes A)\text{vect}(B)$, where $\otimes$ is the Kronecker product of matrices. Using these facts and the properties of the Kronecker product we can argue that

$$\text{vect}(\Delta^+) = Z\text{vect}(\Delta),$$

where

$$\begin{aligned} Z = {} & \{(1-p)I + pP)^* \otimes ((1-p)I + pP)^*\} \\ & + p(1-p)M\{I \otimes (\text{out }(P)\text{out }(P)^*) + \text{out }(P)^* \otimes \text{out }(P)^*\}M \\ & - p(1-p)\{\text{out }(\tilde{P})^* \otimes I + I \otimes \text{out }(\tilde{P})^*\}M. \end{aligned}$$

Then the rate of convergence R will coincide with the absolute value of the dominant reachable eigenvalue of the pair $(Z, \text{vect}(\Delta(0)))$.

Example 2. We apply the previous method for evaluating the rate of convergence for the following matrices:

$$P_1 = \begin{bmatrix} 3/4 & 1/4 & 0 & 0 \\ 1/8 & 1/2 & 3/8 & 0 \\ 0 & 1/8 & 5/8 & 1/4 \\ 1/8 & 1/8 & 0 & 3/4 \end{bmatrix}, \quad P_2 = \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix},$$

$$P_3 = \begin{bmatrix} 1/2 & 1/2 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 1/2 & 1/2 \\ 1/2 & 0 & 0 & 1/2 \end{bmatrix}, \quad P_4 = \begin{bmatrix} 1/3 & 1/3 & 0 & 1/3 \\ 1/3 & 1/3 & 1/3 & 0 \\ 0 & 1/3 & 1/3 & 1/3 \\ 1/3 & 0 & 1/3 & 1/3 \end{bmatrix}.$$

The corresponding rate of convergence is illustrated in Figure 1. We will see in what follows that the same results can be found more easily for the matrices P_2, P_3, P_4.

3.2. The balanced compensation method. In the analysis of this case the following parameters will play a fundamental role:

$$\begin{aligned} \beta_{ih} &:= \mathbb{E}\left[\frac{P_{ih}L_{ih}}{\nu_i}\right] = \mathbb{E}[P_{ih}(0)], \\ \rho_{ihk} &:= \mathbb{E}\left[\frac{P_{ih}L_{ih}P_{ik}L_{ik}}{\nu_i^2}\right] = \mathbb{E}[P_{ih}(0)P_{ik}(0)]. \end{aligned}$$

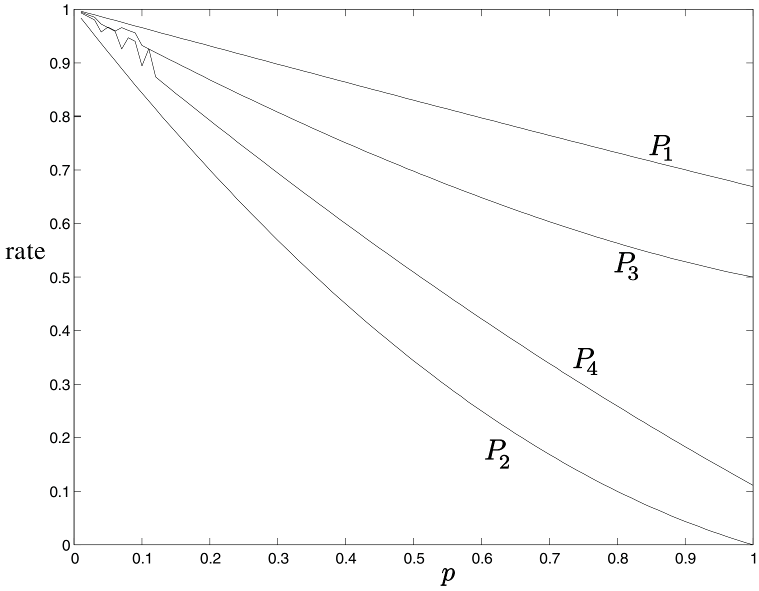


FIG. 1. *The graph of the rate of convergence for the matrices P_1, P_2, P_3, P_4 in Example* 2.

Notice that

$$\beta_{ih} = \sum_{\substack{v\in\{0,1\}^N \\ v_i=1}} \frac{P_{ih}v_h}{\sum_s v_s P_{is}} p^{\mathbf{w}_H(v)-1}(1-p)^{N-\mathbf{w}_H(v)}$$

and

$$\rho_{ihk} = \sum_{\substack{v\in\{0,1\}^N \\ v_i=1}} \frac{P_{ih}v_h P_{ik}v_k}{(\sum_s v_s P_{is})^2} p^{\mathbf{w}_H(v)-1}(1-p)^{N-\mathbf{w}_H(v)},$$

where $\mathbf{w}_H(v)$ is the Hamming weight. Therefore these parameters are polynomial functions of p of degree at most $N-1$. It is clear that $\rho_{ihk} = \rho_{ikh}$. The following lemma presents some other properties.

LEMMA 3.3. *The following relations hold true:*

$$\begin{aligned} &\sum_h \beta_{ih} = 1, \\ &\sum_{hk} \rho_{ihk} = 1, \qquad (13) \\ &\sum_k \rho_{ihk} = \beta_{ih}. \end{aligned}$$

Proof. We prove only the first one. The remaining relations can be proved in a similar way.

$$\begin{aligned} \sum_h \beta_{ih} &= \sum_h \mathbb{E}\left[\frac{P_{ih}L_{ih}}{\nu_i}\right] \\ &= \mathbb{E}\left[\frac{\sum_h P_{ih}L_{ih}}{\nu_i}\right] = 1. \qquad \square \end{aligned}$$

Now define the matrix

$$\bar{\beta} := \{\beta_{ij}\}.$$

Notice that $\bar{\beta} = \mathbb{E}[P(0)]$. By (13) this is a stochastic matrix and its graph coincides with the graph associated with the matrix P. Introduce, moreover, the linear operator $\bar{\rho}$ from the space of diagonal $N \times N$ matrices to the space of symmetric $N \times N$ matrices defined as follows:

$$\bar{\rho}(\text{diag }\{a_1, \dots, a_N\})_{ij} := \sum_k \rho_{kij} a_k.$$

Using these definitions, the relations proposed in the previous lemma can be translated to the following ones:

$$\bar{\beta}\mathbf{1} = \mathbf{1}, \qquad \mathbf{1}^* \bar{\rho}(e_i e_i^*)\mathbf{1} = 1, \qquad \bar{\rho}(e_i e_i^*)\mathbf{1} = \bar{\beta}^* e_i$$

for all $i = 1, \dots, N$, where e_i is the vector with all zeros except for a 1 in the ith position. The second condition is implied by the third and so can be eliminated. Moreover, the third condition is equivalent to the fact that for all diagonal matrices A, we have that

$$\bar{\rho}(A)\mathbf{1} = \bar{\beta}^* A\mathbf{1}.$$

The following result is less immediate to prove.

Lemma 3.4. *If A is a nonnegative diagonal matrix, then $\bar{\rho}(A) - \bar{\beta}^* A\bar{\beta}$ is a positive semidefinite matrix.*

Proof. Notice that

$$\begin{aligned} x^* \bar{\rho}(A)x &= \sum_i \sum_{hk} a_i \rho_{ihk} x_h x_k = \sum_i \sum_{hk} a_i \mathbb{E}\left[\frac{P_{ih}L_{ih}P_{ik}L_{ik}}{\nu_i^2}\right] x_h x_k \\ &= \sum_i a_i \mathbb{E}\left[\sum_{hk} \frac{P_{ih}L_{ih}P_{ik}L_{ik}x_h x_k}{\nu_i^2}\right] \\ &= \sum_i a_i \mathbb{E}\left[\left(\sum_h \frac{P_{ih}L_{ih}x_h}{\nu_i^2}\right)^2\right] \\ &\geq \sum_i a_i \mathbb{E}\left[\sum_h \frac{P_{ih}L_{ih}x_h}{\nu_i^2}\right]^2 = x^* \bar{\beta}^* A\bar{\beta}x. \qquad \square \end{aligned}$$

We are now in a position to present the following result.

Proposition 3.5. *The sequence of matrices $\Delta(t)$ satisfies the recursive relation*

$$\Delta^+ = \bar{\beta}^* \text{ out } (\Delta)\bar{\beta} + \bar{\rho}(\text{diag }(\Delta)). \tag{14}$$

Proof. Notice that

$$\begin{aligned}\mathbb{E}[P(0)^*\Delta P(0)]_{ij} &= \mathbb{E}[Q^*D\Delta DQ]_{ij} = \sum_h\sum_k \mathbb{E}\left[\frac{P_{hi}L_{hi}}{\nu_h}\frac{P_{kj}L_{kj}}{\nu_k}\right]\Delta_{hk}\\ &= \sum_{h\neq k}\mathbb{E}\left[\frac{P_{hi}L_{hi}}{\nu_h}\right]\left[\frac{P_{kj}L_{kj}}{\nu_k}\right]\Delta_{hk} + \sum_k \mathbb{E}\left[\frac{P_{ki}L_{ki}P_{kj}L_{kj}}{\nu_k^2}\right]\Delta_{kk}\\ &= \sum_{k\neq h}\beta_{hi}\Delta_{hk}\beta_{kj} + \sum_k \rho_{kij}\Delta_{kk}\\ &= \{\bar\beta^* \text{ out } (\Delta)\bar\beta + \bar\rho(\text{diag } (\Delta))\}_{ij}.\end{aligned}$$

This easily yields (14). □

Remark 3. Also in this case numerical algorithms can be employed to evaluate the rate of convergence. Introduce, moreover, the $N^2 \times N^2$ matrix T which is zero except in the following entries:

$$T_{(j-1)N+i,(s-1)N+s} = \rho_{sij}.$$

The matrix T is constructed in such a way that, for any diagonal matrix D, we have that

$$\text{vect}(\bar\rho(D)) = T\text{vect}(D).$$

Using the same arguments implemented in the previous remark we can argue that

$$\text{vect}(\Delta^+) = Z\text{vect}(\Delta),$$

where

$$Z = [\bar\beta^* \otimes \bar\beta^*](I - M) + TM,$$

and where the matrix M was defined in (12). Then the rate of convergence R will coincide with the absolute value of the dominant reachable eigenvalue of the pair $(Z, \text{vect}(\Delta(0)))$.

Example 3. We applied the previous method for evaluating the rate of convergence for the same matrices P_1, P_2, P_3, P_4 considered in Example 2. The corresponding rate of convergence is illustrated in Figure 2 and compared with the rates obtained by the biased compensation method. The balanced compensation method outperforms the biased compensation method for all the matrices except for P_3 in which the two methods coincide. We will see in what follows that the same results can be found more easily for the matrices P_2, P_3, P_4.

In what follows we will make further analytical developments assuming the graph $\mathcal{G}$ possesses some more symmetry; more precisely, we will work with Cayley graphs.

4. Cayley matrices over Abelian groups. For graphs possessing symmetries, the theoretical results obtained in the previous section can be refined quite a bit. In this paper we will deal with a special class of symmetric graphs: Abelian Cayley graphs [1].

Let G (with an addition +) be any finite Abelian group of order $|G| = N$, and let S be a subset of G containing zero. The Cayley graph $\mathcal{G}(G, S)$ is the directed graph with vertex set G and arc set

$$\mathcal{E} = \{(g, h) : h - g \in S\}.$$

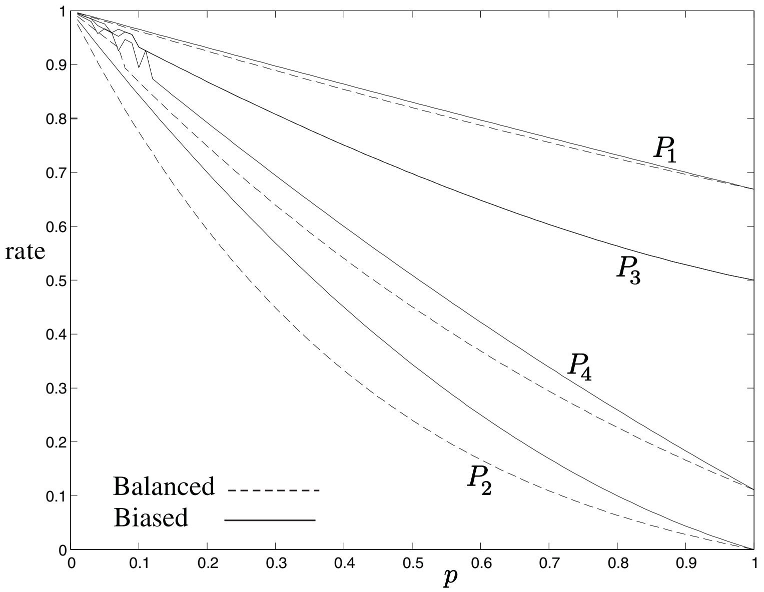


FIG. 2. *The graph of the rate of convergence for the matrices P_1, P_2, P_3, P_4. Biased compensation method rate of convergence is described by the continuous line; balanced compensation method rate of convergence is described by the dashed line.*

Notice that a Cayley graph is always in-regular and out-regular: Both the in-degree and the out-degree of each vertex are equal to $|S|$. Notice also that strong connectivity can be checked algebraically. Indeed, it can be seen that a Cayley graph $\mathcal{G}(G, S)$ is strongly connected if and only if the set S generates the group G, which means that any element in G can be expressed as a finite sum of (not necessarily distinct) elements in S. If S is such that $-S = S$, then the graph obtained is symmetric.

Symmetries can also be introduced on matrices. Let G be any finite Abelian group of order $|G| = N$. A matrix $P \in \mathbb{R}^{G\times G}$ is said to be a Cayley matrix over the group G if

$$P_{i,j} = P_{i+h,j+h} \qquad \forall\ i, j, h \in G\,.$$

It is clear that for a Cayley matrix P there exists a $\pi : G \to \mathbb{R}$ such that $P_{i,j} = \pi(i-j)$. The function π is called the generator of the Cayley matrix P. Notice that, if π and π' are generators of the Cayley matrices P and P', respectively, then $\pi + \pi'$ is the generator of $P + P'$ and $\pi * \pi'$ is the generator of PP', where $(\pi * \pi')(i) := \sum_{j\in G} \pi(j)\pi'(i-j)$ for all $i \in G$. This in particular shows that P and P' commute. It is easy to see that for any Cayley matrix P we have that $P\mathbf{1} = \mathbf{1}$ if and only if $\mathbf{1}^*P = \mathbf{1}^*$. This implies that a Cayley stochastic matrix is automatically doubly stochastic.

4.1. Spectral properties and Fourier analysis of Cayley matrices over Abelian groups. In this subsection we will show that the spectral properties of Cayley matrices over Abelian groups are particularly simple to analyze. We briefly review the theory of Fourier transform over finite Abelian groups (see [19] for a comprehensive treatment of the topic). Let G be a finite Abelian group of order N as above, and let $\mathbb{C}^*$ be the multiplicative group of the nonzero complex numbers. A character on G is a group homomorphism $\chi : G \to \mathbb{C}^*$, namely, a function χ from G to $\mathbb{C}^*$ such that $\chi(g+h) = \chi(g)\chi(h)$ for all $g, h \in G$. Since we have that

$$\chi(g)^N = \chi(Ng) = \chi(0) = 1 \qquad \forall g \in G,$$

it follows that χ takes values on the Nth roots of unity. The character $\chi_0(g) = 1$ for every $g \in G$ is called the trivial character.

The set of all characters of the group G forms an Abelian group with respect to the pointwise multiplication. It is called the character group and denoted by $\hat{G}$. The trivial character χ_0 is the zero of $\hat{G}$. If we consider the vector space $\mathbb{C}^G$ of all functions from G to $\mathbb{C}$ with the canonical Hermitian form

$$\langle f_1, f_2 \rangle = \sum_{g \in G} f_1(g) f_2(g)^* ,$$

then it can be shown that the set $\{N^{-1/2}\chi \mid \chi \in \hat{G}\}$ is an orthonormal basis of $\mathbb{C}^G$.

The Fourier transform of a function $f : G \to \mathbb{C}$ is defined as

$$\hat{f} : \hat{G} \to \mathbb{C} , \quad \hat{f}(\chi) = \sum_{g \in G} \chi(-g) f(g) .$$

Now fix a Cayley matrix P on the Abelian group G generated by the function $\pi_P : G \to \mathbb{R}$. The spectral structure of P is very simple. Namely, it can be shown that the characters $\chi \in \hat{G}$ are eigenvectors of P and so P is diagonalizable. Moreover, the spectrum of P is given by the Fourier transform of the generator π_P of P:

$$\sigma(P) = \{\hat{\pi}_P(\chi) \mid \chi \in \hat{G}\} .$$

Notice that, if A, B are Cayley matrices with Fourier transforms $\hat{\pi}_A(\chi), \hat{\pi}_B(\chi)$, then

$$\hat{\pi}_{A+B}(\chi) = \hat{\pi}_A(\chi) + \hat{\pi}_B(\chi), \qquad \hat{\pi}_{AB}(\chi) = \hat{\pi}_A(\chi)\hat{\pi}_B(\chi).$$

Moreover, observe that, if A is a Cayley matrix, then diag (A) and out (A) are also Cayley and we have

$$\text{diag } (A) = N^{-1} \text{trace } (A) \; I = N^{-1} \sum_{\bar{\chi}} \hat{\pi}_A(\bar{\chi}) \; I .$$

This implies that, for every $\chi \in \hat{G}$,

$$\hat{\pi}_{\text{diag } (A)}(\chi) = N^{-1} \sum_{\bar{\chi} \in \hat{G}} \hat{\pi}_A(\bar{\chi}),$$
$$\hat{\pi}_{\text{out } (A)}(\chi) = \hat{\pi}_A(\chi) - N^{-1} \sum_{\bar{\chi}} \hat{\pi}_A(\bar{\chi}).$$

5. Mean square analysis for Cayley matrices. In this section we will show that when P is a Cayley matrix, the analysis proposed above simplifies considerably. Let G be a finite Abelian of order N, and let P be a Cayley matrix with respect to G. It easily follows from Propositions 3.2 and 3.5 that $\Delta(t)$ are Cayley matrices. This in particular implies that the matrices $\Delta(t)$ admit a common orthonormal basis of eigenvectors. In other words, there exists an $N \times N$ unitary matrix U such that $U^*\Delta(t)U = \tilde{\Delta}(t)$ is diagonal for every t. We can then write

$$\tilde{\Delta}(t+1) = \mathbb{E}[U^* P(0)^* U \tilde{\Delta}(t) U^* P(0) U]. \tag{15}$$

This shows that there exists a linear operator $\tilde{\mathcal{L}}$ such that $\tilde{\Delta}(t+1) = \tilde{\mathcal{L}}(\tilde{\Delta}(t))$ for every t. It is clear that

$$R = \max\{|\lambda| \ : \ \lambda \text{ eigenvalue of } \ \tilde{\mathcal{L}}_{|\tilde{\mathcal{R}}}\},$$

where $\tilde{\mathcal{R}}$ is the reachable subspace of the pair $(\tilde{\mathcal{L}}, \tilde{\Delta}(0))$, where

$$\tilde{\Delta}(0) = U(I - N^{-1}\mathbf{1}\mathbf{1}^*)U^* = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

Notice that (15) is an evolution equation on the eigenvalues of the matrices $\Delta(t)$ which we know to be given by the Fourier transforms $\hat{\pi}_{\Delta(t)}(\chi)$ of the generating function $\pi_{\Delta(t)}(g)$. We will use these representations to express, in a more explicit form, the operator $\tilde{\mathcal{L}}$. As before we will handle the two cases separately.

5.1. The biased compensation method. In the event that P is a Cayley matrix, for the biased compensation method the evolution of the eigenvalues of $\Delta(t)$ is described by the following proposition. First notice that, since P is a Cayley matrix, $\tilde{P}$ and P^*P are also Cayley matrices.

PROPOSITION 5.1. *For all $\chi \in \hat{G}$ we have that*

$$\hat{\pi}_{\Delta(t+1)}(\chi) = A(\chi)\hat{\pi}_{\Delta(t)}(\chi) + B(\chi)N^{-1} \sum_{\bar{\chi} \in \hat{G}} \hat{\pi}_{\Delta(t)}(\bar{\chi}), \tag{16}$$

where

$$A(\chi) = |1 - p + p\hat{\pi}_P(\chi)|^2 \tag{17}$$

and

$$B(\chi) = 2p(1-p)\left\{\hat{\pi}_{\tilde{P}}(\chi_0) - \Re\left[\hat{\pi}_{\tilde{P}}(\chi)\right]\right\}. \tag{18}$$

Proof. We start with formula (9). First notice that, since the matrix diag (Δ) is Cayley and diagonal, it is a scalar multiple of the identity, namely, diag $(\Delta) = xI$. This implies that

$$\begin{aligned}\Delta^+ &= [(1-p)I + pP]^*\Delta[(1-p)I + pP] \\ &\quad + p(1-p)\Big\{\text{diag }[\text{out }(P)\text{out }(P)^* + \text{out }(P)^*\text{out }(P)] - \text{out }(\tilde{P}) \\ &\quad -\text{out }(\tilde{P})^*\Big\}\,\text{diag }(\Delta) \\ &= [(1-p)I + pP]^*\Delta[(1-p)I + pP] + p(1-p)\Big\{\text{diag }[PP^* + P^*P] - \tilde{P} - \tilde{P}^*\Big\}\,xI.\end{aligned}$$

Notice now that

$$\text{diag }[PP^* + P^*P] = 2\hat{\pi}_{\tilde{P}}(\chi_0)\ I$$

and that

$$x = N^{-1} \sum_{\bar{\chi} \in \hat{G}} \hat{\pi}_{\Delta}(\bar{\chi}).$$

These facts yield (16). □

Remark 4. Notice that $\tilde{P}$ is an irreducible nonnegative Cayley matrix and so $\hat{\pi}_{\tilde{P}}(\chi_0)$ is its spectral radius. This implies that $B(\chi) \geq 0$ and that $B(\chi) = 0$ if and only if $\chi = \chi_0$.

The linear dynamic system described in (16) can finally be rewritten in a more compact way as follows. Enumerate in some way the characters of G, $\hat{G} = \{\chi_0, \chi_1, \dots, \chi_{N-1}\}$, and define the column vector in $\mathbb{R}^N$ as

$$\Pi(t) := \begin{bmatrix} \hat{\pi}_{\Delta(t)}(\chi_0) \\ \vdots \\ \hat{\pi}_{\Delta(t)}(\chi_{N-1}) \end{bmatrix}. \tag{19}$$

Define, moreover, the column vector B in $\mathbb{R}^N$ such that for all $i = 0, 1, \dots, N-1$, we have that $B_i := B(\chi_i)$ and the diagonal matrix A such that for all $i = 0, 1, \dots, N-1$, we let $A_{ii} := A(\chi_i)$. Then we can write the linear dynamic system (16) as follows:

$$\Pi(t+1) = \left(A + N^{-1} B \mathbf{1}^*\right) \Pi(t).$$

Notice that both A and B depend on the probability p and so in some cases we will write $A(p)$ and $B(p)$ to make this dependence evident. Notice that $A_{ii}(p) < 1$ if $p > 0$, while $A_{ii}(0) = 1$ for all i. Moreover, we have that $B_0(p) = 0$ for all i and $0 < B_i(p) < 1$ if $i \neq 0$ and $0 < p < 1$, while $B_i(0) = B_i(1) = 0$.

We have the following result.

Proposition 5.2. *We have the following properties:*

(a) *The matrix $A + N^{-1} B \mathbf{1}^*$ has nonnegative entries.*

(b) *It has the structure*

$$A + N^{-1} B \mathbf{1}^* = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ X_{21} & & X_{22} & \end{pmatrix}, \tag{20}$$

where $X_{21} \in \mathbb{R}^{1 \times (N-1)}$ and $X_{22} \in \mathbb{R}^{(N-1) \times (N-1)}$ have nonnegative entries.

(c) $R = \max\{|\lambda| \ : \ \lambda \text{ eigenvalue of } \ X_{22}\} \geq \max\{A_{ii} : 1 = 1, \dots, N-1\}$.

(d) *The eigenvector of $A + N^{-1} B \mathbf{1}^*$ relative to the eigenvalue* 1 *has a nonzero first component.*

Proof. (a) follows from the previous remark.

(b) can be proven by inspection.

(c) Notice that $\Pi(0) = (0, 1, \dots, 1)^*$. Because of (a), this shows that the reachability subspace $\tilde{\mathcal{R}}$ of the pair $(X, \Pi(0))$ has the structure $\tilde{\mathcal{R}} = \{0\} \times \tilde{\mathcal{R}}_2$, where $\tilde{\mathcal{R}}_2$ is the reachability subspace of $(X_{22}, \mathbf{1})$. Since X_{22} has nonnegative entries, its spectral radius is achieved by a nonnegative eigenvalue λ with a corresponding nonnegative eigenvector $\bar{w}$ [7, p. 66]. Clearly, we can write $\mathbf{1} = a\bar{w} + w'$ for some $a > 0$ and another nonnegative vector w'. Hence,

$$X_{22}^t \mathbf{1} = a\lambda^t \bar{w} + X_{22}^t w' \geq a\lambda^t \bar{w}\,.$$

From this it immediately follows that $R \geq \lambda$. This clearly proves the equality in (c). Finally, the fact that $R \geq \max\{A_{ii} : 1 = 1, \dots, N-1\}$ follows from the fact that B has nonnegative entries [10, Corollary 8.1.19].

(d) Finally, consider any eigenvector $w \in \mathbb{R}^N$ of $A+N^{-1}B\mathbf{1}^*$ relative to the eigenvalue 1. If the first component w_0 of w were zero, then the vector $(w_1, \dots, w_{N-1})^*$ would be an eigenvector of X_{22} relative to the eigenvalue 1. This could not be possible, however, since we know that 1 is an eigenvalue of $A + N^{-1}B\mathbf{1}^*$ with algebraic multiplicity equal to 1. □

Proposition 5.2 reduces the computation of R to the computation of the second dominant eigenvalue of the $N \times N$ matrix (20). An explicit expression for the characteristic polynomial of (20) can be obtained through the following lemma.

LEMMA 5.3.

$$\det\left(A + N^{-1}B\mathbf{1}^*\right) = \prod_{j=0}^{N-1} A_{jj} + N^{-1} \sum_{i=0}^{N-1} B_i \prod_{\substack{j=0\\ j\neq i}}^{N-1} A_{jj}$$

Proof. Notice that

$$\begin{aligned}&\det\left(A + N^{-1}B\mathbf{1}^*\right)\\ &\quad= \det(A)\det\left(I + N^{-1}B\mathbf{1}^*A^{-1}\right) = \det(A)\left(1 + N^{-1}\mathbf{1}^*A^{-1}B\right). \qquad \square\end{aligned}$$

From this lemma we can argue that

$$\begin{aligned}F(z,p) &:= \det\left(zI - A(p) - N^{-1}B(p)\mathbf{1}^*\right)\\ &= \prod_{j=0}^{N-1}(z - A_{jj}(p)) - N^{-1}\sum_{i=0}^{N-1} B_i(p) \prod_{\substack{j=0\\ j\neq i}}^{N-1}(z - A_{jj}(p)).\end{aligned}$$

The polynomial $F(z,p)$ has degree N in z and degree $2n$ in p. The stability analysis of this polynomial can be in general quite complicated. We will investigate this problem through some examples.

We start with a couple of examples in which the eigenvalues can be determined exactly and also some natural optimization design can be carried out.

Example 4. Consider the matrix $P = (1-k)I + \frac{k}{N}\mathbf{1}\mathbf{1}^*$. It is clear that the matrix P is in this case a Cayley matrix over the group $\mathbb{Z}_N$ and with $S = \mathbb{Z}_N$. After some computation we can find that

$$A_{ii}(k,p) = \begin{cases} 1 & \text{if } i = 0,\\ (1-kp)^2 & \text{if } i \neq 0\end{cases} \tag{21}$$

and

$$B_i(k,p) = \begin{cases} 0 & \text{if } i = 0,\\ \frac{2p(1-p)k^2}{N} & \text{if } i \neq 0,\end{cases} \tag{22}$$

and so the eigenvalues are

$$\begin{aligned}\bar{z}_0(k,p) &= 1,\\ \bar{z}_1(k,p) &= (1-kp)^2 + 2p(1-p)k^2\frac{N-1}{N^2},\\ \bar{z}_i(k,p) &= (1-kp)^2 \qquad i = 2, \dots, N-1.\end{aligned}$$

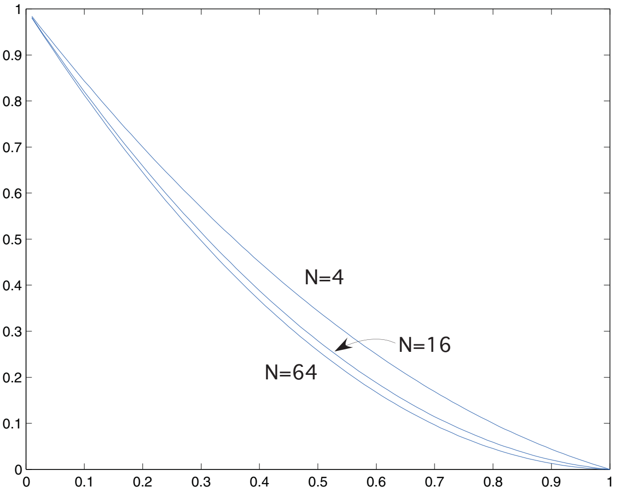


FIG. 3. *The graph of the rate of convergence in Example* 4 *for* $N = 4, 16, 64$.

The rate of convergence to the consensus is determined by the eigenvalue $\bar{z}_1(k, p)$. In this case the optimal k yielding the fastest convergence can be computed analytically. Indeed, it can be seen that

$$k = \frac{N^2}{N^2 p + 2(1-p)(N-1)}.$$

For large N we have that $k \simeq 1/p$. In Figure 3 we show the graph of the rate of convergence as a function of the probability p for $N = 4, 16, 64$ when $k = 1$. It can be shown that the graph relative to the case $N = 4$ coincides up to numerical errors with the one obtained in Example 2 for the matrix P_2.

Example 5. Consider the case in which the group is $\mathbb{Z}_N$ and $S = \{0, 1\}$. Consider a matrix P with generator $\pi_P(0) = 1 - k, \pi_P(1) = k$, and $\pi_P(g) = 0$ for all $g \neq 0, 1$. In this case we have that

$$\hat{\pi}_P(\chi_i) = 1 - k + ke^{j\frac{2\pi}{N}i},$$
$$\hat{\pi}_{\tilde{P}}(\chi_i) = (1-k)^2 + k^2 e^{j\frac{2\pi}{N}i}.$$

From this we can argue that

$$A_{ii}(p) = 1 - 2pk(1-pk)\left(1 - \cos\left(\frac{2\pi}{N}i\right)\right), \tag{23}$$

$$B_i(p) = 2p(1-p)k^2\left(1 - \cos\left(\frac{2\pi}{N}i\right)\right). \tag{24}$$

With fixed probability p one can find the optimal k yielding the fastest convergence. We did this numerically for $N = 5, 10, 20$. The graph showing the optimal k as a function of p is illustrated in Figure 4. In Figure 5 we show the graph of the rate of

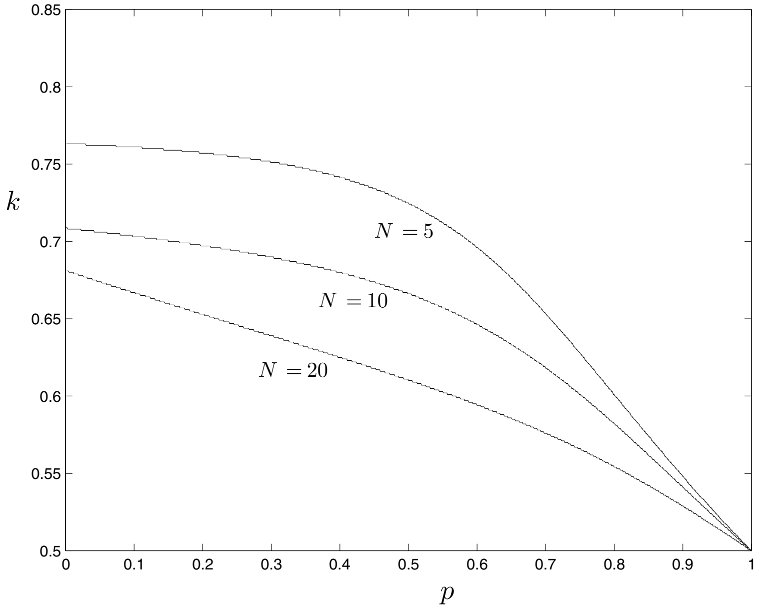


FIG. 4. *The graph of the optimal k as a function of the probability p in Example* 5 *for $N = 5, 10, 20$.*

convergence as a function of the probability p for $N = 2, 4, 8$ when $k = 1/2$. It can be shown that the graph relative to the case $N = 4$ coincides up to numerical errors with the one obtained in Examples 2 and 3 for the matrix P_3.

The next example relates instead to the hypercube graph. We present only the analytical computation of the eigenvalues.

Example 6. Consider the case in which the group is $\mathbb{Z}_2^n$ and

$$S = \{0, e_1, \dots, e_n\},$$

where e_i is the vector with all zeros except for a 1 in the ith position. Let E be the adjacency matrix of the graph defined in this way and consider the matrix $P := \frac{1}{n+1}E$. This means that given $u, v \in \mathbb{Z}_2^n$ we have that

$$P_{u,v} = \begin{cases} \frac{1}{n+1} & \text{if } u + v \in S, \\ 0 & \text{otherwise.} \end{cases}$$

Notice that, in this case, we have that $\tilde{P} = \frac{1}{n+1}P$. It can be shown that for all $v \in \mathbb{Z}_2^n$ we have that

$$\hat{\pi}_P(v) = 1 - \frac{2}{n+1}\mathbf{w}_H(v),$$

where $\mathbf{w}_H(v)$ is the Hamming weight of v, namely, the number of 1's. From this we can argue that

$$A_v = \left(1 - \frac{2p}{n+1}\mathbf{w}_H(v)\right)^2,$$

$$B_v = \frac{4p(1-p)}{(n+1)^2}\mathbf{w}_H(v).$$

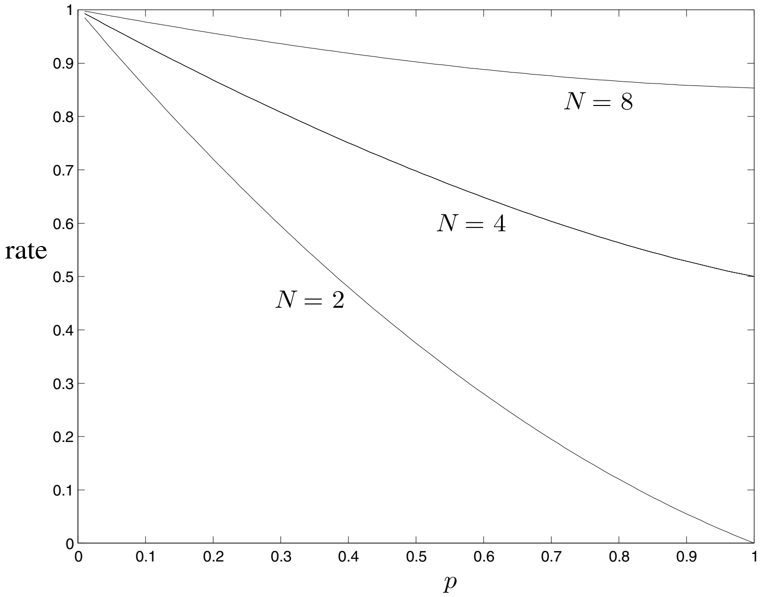


FIG. 5. *The graph of the rate of convergence in Example* 5 *for* $N = 2, 4, 8$.

From this it follows that

$$F(z,p) = \prod_{h=0}^{n} (z - A_h)^{\binom{n}{h}-1} \left\{ \prod_{h=0}^{n} (z - A_h) - N^{-1} \sum_{k=0}^{n} \binom{n}{k} B_k \prod_{\substack{h=0 \\ h\neq k}}^{n} (z - A_h) \right\},$$

where, for $h = 0, 1, \ldots, n$, we let

$$A_h = \left(1 - \frac{2p}{n+1} h\right)^2,$$
$$B_h = \frac{4p(1-p)}{(n+1)^2} h.$$

This implies that $N - n$ eigenvalues coincide with $A_h(p)$, $h = 0, 1, \ldots, n$, while the remaining n are the roots of

$$\prod_{h=0}^{n} (z - A_h) - N^{-1} \sum_{k=0}^{n} \binom{n}{k} B_k \prod_{\substack{h=0 \\ h\neq k}}^{n} (z - A_h).$$

Figure 7 shows the graph of the rate of convergence as a function of the probability p for $n = 2, 4$. It can be shown that the graph relative to the case $n = 2$ coincides up to numerical errors with the one obtained in Example 2 for the matrix P_4.

Now we present an example where instead only numerical results can be obtained.

5.2. The balanced compensation method. First notice that, if P is a Cayley matrix, we have the following result.

LEMMA 5.4. *If P is a Cayley matrix, then $\bar{\beta}, \bar{\rho}(I)$ are also Cayley matrices.*

Proof. Notice that

$$\begin{aligned}\beta_{i+l,h+l} &= \sum_{\substack{v\in\{0,1\}^N\\ v_{i+l}=1}} \frac{P_{i+l,h+l}v_{h+l}}{\sum_s v_s P_{i+l,s}} p^{\mathbf{w}_H(v)-1}(1-p)^{N-\mathbf{w}_H(v)}\\ &= \sum_{\substack{v\in\{0,1\}^N\\ v_{i+l}=1}} \frac{P_{i,h}v_{h+l}}{\sum_s v_s P_{i,s-l}} p^{\mathbf{w}_H(v)-1}(1-p)^{N-\mathbf{w}_H(v)}.\end{aligned}$$

We now define for any $v\in\{0,1\}^N$ a $u\in\{0,1\}^N$ such that $v_s = u_{s-l}$. Then

$$\begin{aligned}\beta_{i+l,h+l} &= \sum_{\substack{u\in\{0,1\}^N\\ u_i=1}} \frac{P_{i,h}u_h}{\sum_s u_{s-l} P_{i,s-l}} p^{\mathbf{w}_H(u)-1}(1-p)^{N-\mathbf{w}_H(u)}\\ &= \sum_{\substack{u\in\{0,1\}^N\\ u_i=1}} \frac{P_{i,h}u_h}{\sum_s u_s P_{i,s}} p^{\mathbf{w}_H(u)-1}(1-p)^{N-\mathbf{w}_H(u)} = \beta_{i,h}.\end{aligned}$$

In a similar way we can prove that $\rho_{k,i,j} = \rho_{k+l,i+l,j+l}$. From this it follows that $\bar\rho(I)$ is a Cayley matrix. □

In this case we have the following proposition.

PROPOSITION 5.5. *For all $\chi\in\hat G$ we have that*

$$\hat\pi_{\Delta(t+1)}(\chi) = A(\chi)\hat\pi_{\Delta(t)}(\chi) + B(\chi)N^{-1}\sum_{\bar\chi\in\hat G}\hat\pi_{\Delta(t)}(\bar\chi), \tag{25}$$

where

$$A(\chi) = \left|\hat\pi_{\bar\beta}(\chi)\right|^2 \tag{26}$$

and

$$B(\chi) = \hat\pi_{\bar\rho(I)}(\chi) - \left|\hat\pi_{\bar\beta}(\chi)\right|^2 . \tag{27}$$

Proof. We start from the equation

$$\Delta^+ = \bar\beta^* \text{out } (\Delta)\bar\beta + \bar\rho(\text{diag } (\Delta)).$$

Notice now that diag (Δ) is a multiple of the identity, namely diag $(\Delta) = xI$. This implies that

$$\Delta^+ = \bar\beta^*\Delta\bar\beta + \{\bar\rho(I) - \bar\beta^*\bar\beta\}x.$$

Considering the fact that

$$x := N^{-1}\sum_{\bar\chi}\hat\pi_\Delta(\bar\chi),$$

we obtain the thesis. □

As in the previous case, using the notation (19), we can rewrite (25) as

$$\Pi(t+1) = (A + N^{-1}B\mathbf{1}^*)\Pi(t),$$

where B is the column vector in $\mathbb{R}^N$ such that for all $i = 0, 1, \ldots, N-1$, we have that $B_i := B(\chi_i)$ and A is the diagonal matrix such that $A_{ii}(\chi) := A(\chi_i)$. As in the previous case we will use the notation $A(p)$ and $B(p)$ whenever we want to underline the dependence on p.

Notice that, as observed above, when $p > 0$, the matrix $\bar{\beta}$ is an irreducible stochastic matrix. This implies that $A_{ii}(p) < 1$ if $p > 0$. On the other hand, since when $p = 0$ we have $\bar{\beta} = \bar{\rho}(I) = I$, then $A_{ii}(0) = 1$ and $B_i(0) = 0$ for all i. Finally, when $p = 1$ we have that $\bar{\beta} = P$ and $\bar{\rho}(I) = P^*P$, and so $B_i(1) = 0$.

From Lemma 3.3 we can argue that $\hat{\pi}_{\bar{\rho}(I)}(\chi_0) = \hat{\pi}_{\bar{\beta}}(\chi_0) = 1$ and so $B_0(p) = 0$ for all p. Using Lemma 3.4 it can be shown that Proposition 5.2 still holds and as above we can argue that the eigenvalues of $A(p) + N^{-1}B(p)\mathbf{1}^*$ coincide with the roots of the polynomial

$$F(z,p) = \prod_{j=0}^{N-1}(z - A_{jj}(p)) - N^{-1}\sum_{i=0}^{N-1} B_i(p) \prod_{\substack{j=0\\ j\neq i}}^{N-1}(z - A_{jj}(p)).$$

In some cases some further simplifications can be introduced. Consider a Cayley graph $\mathcal{G}$. Since each node of $\mathcal{G}$ has exactly the same number n (excluding self loops) of incoming edges and outgoing edges, we can introduce a Cayley matrix $\bar{P}$ compatible with $\mathcal{G}$ by letting

$$\bar{P}_{ij} = \begin{cases} 1/n & \text{if } (j,i) \text{ is an edge of the graph,} \\ 0 & \text{otherwise.} \end{cases}$$

Moreover, let

$$P := (1-k)I + k\bar{P}. \tag{28}$$

In this way we obtained a family of Cayley matrices P compatible with the graph $\mathcal{G}$. In this case the parameters β_{ih}, ρ_{ihk} become simpler to evaluate. Indeed, let E be the adjacency matrix of the graph and $H := E - I$. Moreover, let b_k be a binomial random variable, namely, a random variable taking value on the nonnegative integers with law

$$\mathbb{P}[b_k = i] = \binom{k}{i} p^i (1-p)^{k-i}, \quad i = 0, 1, \ldots, k.$$

After some simple but lengthy calculations it can be shown that

$$\begin{array}{ll} \beta_{ii} = \alpha & \forall\, i, \\ \beta_{ih} = \beta H_{ih} & \forall\, i, h \text{ such that } i \neq h, \\ \rho_{iii} = \gamma, & \\ \rho_{iih} = \rho_{ihi} = \delta H_{ih} & \forall\, i, h \text{ such that } i \neq h, \\ \rho_{ihh} = \xi H_{ih} & \forall\, i, h \text{ such that } i \neq h, \\ \rho_{ihk} = \rho H_{ih} H_{ik} & \forall\, i, h, k \text{ that are different from each other,} \end{array}$$

where
(29)

$$\alpha := \mathbb{E}\left[\frac{n-nk+k}{n-nk+k+kb_{n-1}}\right], \qquad \beta := p\mathbb{E}\left[\frac{k}{n-nk+2k+kb_{n-2}}\right],$$

$$\gamma := \mathbb{E}\left[\frac{(n-nk+k)^2}{(n-nk+k+kb_{n-1})^2}\right], \qquad \delta := p\mathbb{E}\left[\frac{(n-nk+k)k}{(n-nk+2k+kb_{n-2})^2}\right],$$

$$\xi := p\mathbb{E}\left[\frac{k^2}{(n-nk+2k+kb_{n-2})^2}\right], \qquad \rho := p^2\mathbb{E}\left[\frac{k^2}{(n-nk+3k+kb_{n-3})^2}\right].$$

These parameters depend on k, n, and p. The relations (13) become

$$\begin{aligned}\alpha + \beta(n-1) &= 1,\\ \gamma + \delta(n-1) &= \alpha,\\ (n-2)\rho &= \beta - \xi - \delta.\end{aligned}$$

It is clear that $\bar{\beta} = \alpha I + \beta H$. Moreover, after some computations, it can be shown that, if D is any diagonal matrix, then in this case

$$\bar{\rho}(D) = \rho H^* D H + (\xi - \rho)\text{diag}\,(H^* D H) + \gamma D + \delta(H^* D + DH). \tag{30}$$

Under these assumptions we can write

$$A(\chi) = |\alpha + \beta \hat{\pi}_H(\chi)|^2$$

and

$$B(\chi) = \rho\,|\hat{\pi}_H(\chi)|^2 + (\xi - \rho)(n-1) + \gamma + 2\delta\,\Re\left[\hat{\pi}_H(\chi)\right] - |\alpha + \beta\hat{\pi}_H(\chi)|^2.$$

We now want to compare the two compensation methods proposed here through the examples presented previously. Notice that in Example 5 the two compensation methods coincide.

Example 7. Now consider the matrix $P = (1-k)I + N^{-1}\mathbf{1}\mathbf{1}^*$ considered in Example 4. After some computation we can find that, for $i = 1, \ldots, N-1$,

$$A_{ii}(k,p) = (1 - N\beta)^2, \quad B_i(k,p) = (1-N\beta)N\beta + N(\xi - \delta) \tag{31}$$

and so the eigenvalues are

$$\begin{aligned}\bar{z}_0(k,p) &= 1,\\ \bar{z}_1(k,p) &= (1-N\beta)(1-\beta) + (N-1)(\xi - \delta),\\ \bar{z}_i(k,p) &= (1-\beta N)^2, \qquad i = 2, \ldots, N-1.\end{aligned}$$

The rate of convergence to the consensus is determined by $\bar{z}_1(k,p)$. Figure 6 shows the graph of the dominant eigenvalue $\bar{z}_1(1,p)$ as a function of the probability p for $N = 4, 16, 64$ when $k = 1$. (We are not making any optimization in this case.) It can be shown that the graph relative to the case $N = 4$ coincides up to numerical errors with the one obtained in Example 2 for the matrix P_2.

Example 8. Consider the same matrix P introduced in Example 6. It can be shown that for all $v \in \mathbb{Z}_2^n$ we have that

$$\begin{aligned}A_v &= (1 - 2\beta\mathbf{w}_H(v))^2,\\ b_v &= 4\mathbf{w}_H(v)(\delta - \rho + \mathbf{w}_H(v)(\rho - \beta^2)),\end{aligned}$$

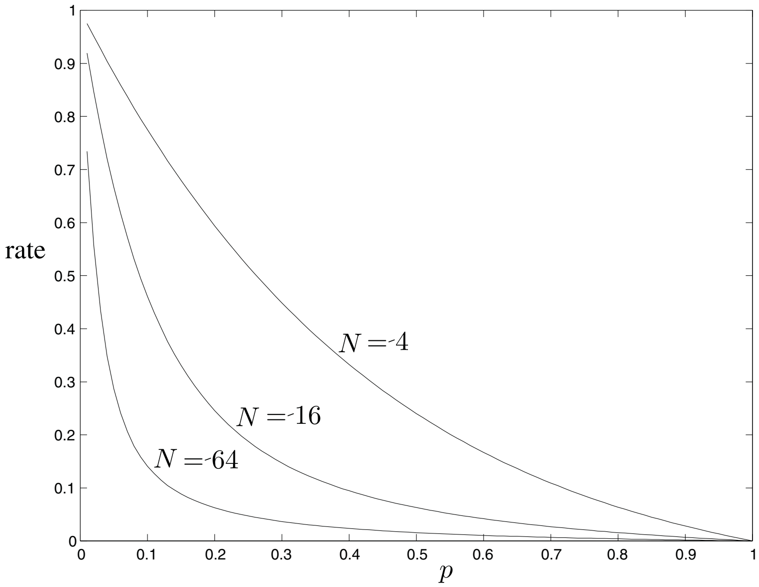


FIG. 6. *The graph of the rate of convergence in Example* 7 *for* $N = 4, 16, 64$.

where $\mathbf{w}_H(v)$ is the Hamming weight of v, namely, the number of 1's. From this it follows that

$$F(z,p) = \prod_{h=0}^{n} (z - A_h)^{\binom{n}{h}-1} \left\{ \prod_{h=0}^{n} (z - A_h) - N^{-1} \sum_{k=0}^{n} \binom{n}{k} b_k \prod_{\substack{h=0 \\ h \neq k}}^{n} (z - A_h) \right\},$$

where, for $h = 0, 1, \ldots, n$, we let

$$\begin{aligned} A_h &= (1 - 2\beta h)^2, \\ b_h &= 4h(\delta - \rho + h(\rho - \beta^2)). \end{aligned}$$

This implies that $N - n$ eigenvalues coincide with $A_h(p)$, $h = 0, 1, \ldots, n$ while the remaining n are the roots of

$$\prod_{h=0}^{n} (z - A_h) - N^{-1} \sum_{k=0}^{n} \binom{n}{k} b_k \prod_{\substack{h=0 \\ h \neq k}}^{n} (z - A_h)$$

and can be estimated when $p \simeq 1$ by the method proposed above. Figure 7 shows the graph of the rate of convergence as a function of the probability p for $N = 2, 4$ in the case of the biased compensation and the balanced compensation methods. It can be shown that the graph relative to the case $n = 2$ coincides up to numerical errors with the one obtained in Example 2 for the matrix P_4.

6. Average consensus. Even if the original algorithm was chosen to solve the average consensus problem, in general the perturbed solutions due to packet drops will no longer satisfy this property. In this section we will show how to estimate the distance of the consensus point from the average of the initial conditions. From

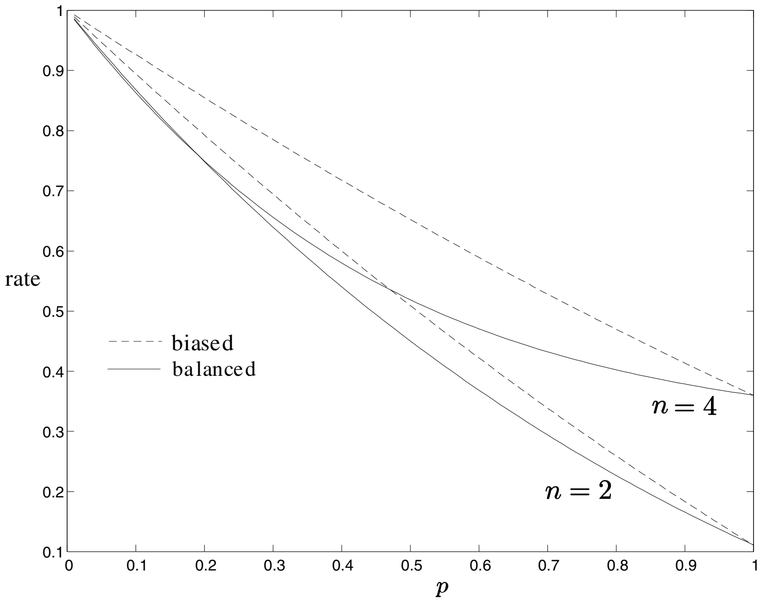


FIG. 7. *The graph of the rate of convergence in Examples* 6 *and* 8 *for* $n = 2, 4$ *in the case of the biased compensation method and the balanced compensation method.*

now on we will assume that the matrix P is doubly stochastic so that $P\mathbf{1} = \mathbf{1}$ and $\mathbf{1}^*P = \mathbf{1}^*$. Consider $x_A(t)$ as defined in (5) and let

$$\begin{aligned} D := & \sup_{||x(0)||\leq 1} \mathbb{E}[|x_A(\infty) - x_A(0)|^2] = \sup_{||x(0)||\leq 1} \mathbb{E}[|(v^* - N^{-1}\mathbf{1}^*)x(0)|^2] \\ = & \sup_{||x(0)||\leq 1} x(0)^*\mathbb{E}[(v - N^{-1}\mathbf{1})(v - N^{-1}\mathbf{1})^*]x(0) \\ = & \max\{|\lambda| \ : \ \lambda \text{ eigenvalue of } \mathbb{E}[(v - N^{-1}\mathbf{1})(v - N^{-1}\mathbf{1})^*]\}\,. \end{aligned}$$

Notice that D is expressed in terms of the random vector v which in general may not be explicitly available. A further step, however, allows us to write

$$\mathbb{E}[(v - N^{-1}\mathbf{1})(v - N^{-1}\mathbf{1})^*] = \mathbb{E}[vv^*] - N^{-1}\mathbb{E}[v]\mathbf{1}^* - \mathbf{1}N^{-1}\mathbb{E}[v]^* + N^{-2}\mathbf{1}\mathbf{1}^*. \tag{32}$$

We now recall that $\mathbb{E}[vv^*]$ is the dominant eigenvector of the positive operator $\mathcal{L}$ and can thus be computed using standard techniques. As far as $\mathbb{E}[v]$ is concerned, notice that, since $x(t) \to v^*x(0)\mathbf{1}$ almost surely, it follows that

$$\mathbb{E}[x(t)] = \mathbb{E}[P(0)]^t x(0) \to \mathbb{E}[v]^*x(0)\mathbf{1}.$$

Since $\mathbb{E}[P(0)]$ is an aperiodic stochastic matrix, it follows that $\mathbb{E}[v]$ coincides with the the dominant left eigenvector of $\mathbb{E}[P(0)]$ and thus it is computable using standard techniques.

We now start to analyze the Cayley setting for which more precise results can be obtained. First, it can be checked that, for both the biased and the balanced compensation methods, the matrix $\overline{P}$ is Cayley. As a consequence, in this case we have $\mathbb{E}[v] = N^{-1}\mathbf{1}$ and hence

$$\mathbb{E}[(v - N^{-1}\mathbf{1})(v - N^{-1}\mathbf{1})^*] = \mathbb{E}[vv^*] - N^{-2}\mathbf{1}\mathbf{1}^*.$$

What remains to be computed is the normalized dominant eigenvector of $\mathcal{L}$ or, equivalently, the normalized dominant eigenvector of the matrix $A + N^{-1}B\mathbf{1}^*$ introduced above, where the matrix A is diagonal such that $A_{ii} = A(\chi_i)$ and B is a column vector such that $B_i = B(\chi_i)$. The quantities $A(\chi)$ and $B(\chi)$ have been defined in (17) and (18) for the biased case, and in (26) and (27) for the balanced case. Notice that, in both cases, we have that

$$\begin{aligned} A(\chi_0) = 1\,,\ 0 \leq A(\chi_i) < 1 \ \ \forall i = 1, \dots, N-1, \\ B(\chi_0) = 0\,,\ B(\chi_i) \geq 0 \ \ \forall i = 1, \dots, N-1\,. \end{aligned}$$

We have the following result.

LEMMA 6.1. *The vector $w \in \mathbb{R}^N$ with components*

$$w_0 = 1\,,\ \ w_h = \frac{N^{-1}}{1 - N^{-1}\sum_{i=1}^{N-1} \frac{B_i}{1-A_{ii}}} \frac{B_h}{1 - A_{hh}} \ \ \forall h = 1, \dots, N-1$$

is an eigenvector of $(A + N^{-1}B\mathbf{1}^*)$ relative to the eigenvalue 1.

Proof. Notice first that, since $A + N^{-1}B\mathbf{1}^*$ is a nonnegative matrix, there exists an eigenvector $w \in \mathbb{R}^N$ of $A + N^{-1}B\mathbf{1}^*$ relative to the eigenvalue 1 with nonnegative entries. By Proposition 5.2 we can argue that the first component w_0 of w must be positive. Notice now that the relation $(A+N^{-1}B\mathbf{1}^*)w = w$ is equivalent to the $N-1$ linear conditions

$$(1 - A_{hh})w_h = N^{-1}B_h\mathbf{1}^*w\,,\ h = 1, \dots, N-1\,. \tag{33}$$

Since, as noticed above, $A_{hh} < 1$, we have that

$$w = \left(w_0, \lambda\frac{B_1}{1 - A_{11}}, \dots, \lambda\frac{B_{N-1}}{1 - A_{N-1\,N-1}}\right)^*, \tag{34}$$

where $\lambda = N^{-1}\mathbf{1}^*w$. This implies that

$$w_0 + \lambda\sum_{i=1}^{N-1} \frac{B_i}{1 - A_{ii}} = N\lambda,$$

which is equivalent to

$$\left(1 - N^{-1}\sum_{i=1}^{N-1} \frac{B_i}{1 - A_{ii}}\right)\lambda = N^{-1}w_0.$$

Finally, notice that, since $w_0 > 0$,

$$1 > \frac{N^{-1}\sum_{i=1}^{N-1} w_i}{\lambda} = N^{-1}\sum_{i=1}^{N-1} \frac{B_i}{1 - A_{ii}},$$

which implies that $1 - N^{-1}\sum_{i=1}^{N-1} \frac{B_i}{1-A_{ii}} < 1$ and so, by taking $w_0 = 1$, we obtain the thesis. □

Notice that, if we go back to the matrix form, the corresponding eigenmatrix of $\mathcal{L}$ is given by

$$W = N^{-1}\sum_{i=0}^{N-1} w_i\chi_i\chi_i^*\,.$$

To find the right normalization constant, notice that $\mathbf{1}^* W \mathbf{1} = N$. This implies that

$$\mathbb{E}[vv^*] = N^{-2} \sum_{i=0}^{N-1} w_i \chi_i \chi_i^* \,. \tag{35}$$

Notice that, since $\mathbb{E}[vv^*]$ is positive semidefinite, surely all $w_i \geq 0$. We can now state the following proposition.

PROPOSITION 6.2. *Assume P to be a Cayley matrix. Then, for both the biased and the balanced compensation methods, we have that*

$$D = \frac{N^{-2}}{1 - N^{-1} \sum\limits_{i=1}^{N-1} \frac{B_i}{1-A_{ii}}} \max_{h=1,\dots,N-1} \left\{ \frac{B_h}{1 - A_{hh}} \right\}.$$

Proof. Notice that, from (32), we obtain

$$\begin{aligned} &\mathbb{E}[(v - N^{-1}\mathbf{1})(v - N^{-1}\mathbf{1})^*] \\ &= N^{-2}\mathbf{1}\mathbf{1}^* + N^{-2} \sum_{i=1}^{N-1} w_i \chi_i \chi_i^* - 2N^{-2}\mathbf{1}\mathbf{1}^* + N^{-2}\mathbf{1}\mathbf{1}^* = N^{-2} \sum_{i=1}^{N-1} w_i \chi_i \chi_i^* \,. \end{aligned}$$

Notice now that

$$D = \max\{N^{-1} w_i \mid i = 1, \dots, N-1\} \,.$$

This proves the result. □

Let us make explicit computations in the examples considered above.

Example 9. Consider the matrix $P = (1-k)I + \frac{k}{N}\mathbf{1}\mathbf{1}^*$ introduced in Examples 4 and 7. In the biased compensation method, using computation (21) and (22), we obtain that for any $h \neq 0$,

$$\frac{B_h}{1 - A_{hh}} = \frac{2p(1-p)k^2}{N} \frac{1}{1 - (1 - p + p(1-k))^2} = N^{-1} \frac{2(1-p)k}{2 - pk} \,.$$

Hence,

$$D = \frac{N^{-2}}{1 - N^{-2}(N-1)\frac{2(1-p)k}{2-pk}} N^{-1} \frac{2(1-p)k}{2-pk} = N^{-3} \frac{2(1-p)k}{2 - pk - 2N^{-2}(N-1)(1-p)k} \,.$$

As expected for $p \to 1$ we have that $D \to 0$. More interestingly, note also that for $N \to +\infty$, we have that $D \to 0$ as N^{-3}.

Consider now the balanced case. We limit the analysis to the optimal case in which we let $k = 1$. Using computation (31), we obtain that for any $h \neq 0$,

$$\frac{B_h}{1 - A_{hh}} = \frac{(1 - N\beta)N\beta}{1 - (1 - N\beta)^2} = \frac{1 - N\beta}{2 - N\beta} \,.$$

Hence,

$$D = \frac{N^{-2}}{1 - N^{-1}(N-1)\frac{1-N\beta}{2-N\beta}} \frac{1 - N\beta}{2 - N\beta} = N^{-2} \frac{1 - N\beta}{1 + N^{-1}(1 - N\beta)} \,.$$

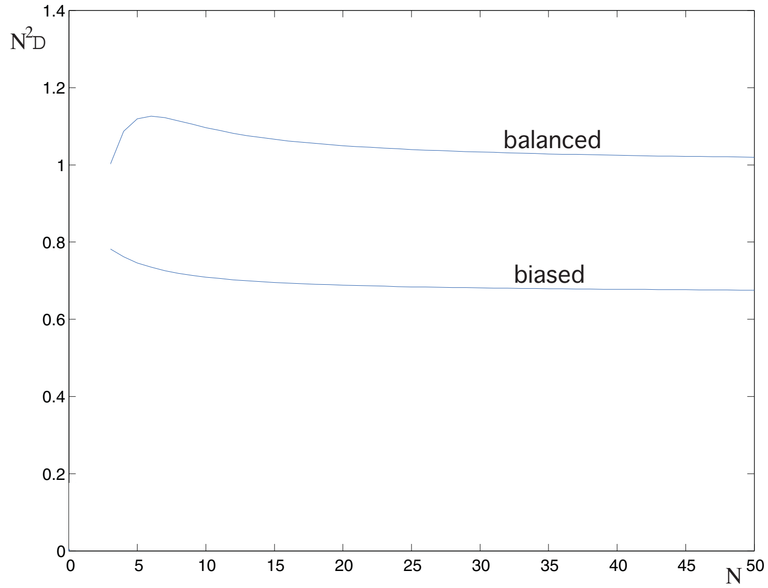


FIG. 8. *The graphs of N^3D as a function of N in the biased and balanced cases, both assuming that $k = 1$ and $p = 1/2$ as in Example* 9.

Even in this case, for $p \to 1$ we have that $D \to 0$ since it is easy to see that $N\beta \to 1$. Also the convergence to 0 for $N \to \infty$ is maintained. In Figure 8 we plot the graphs showing N^3D as a function of N in the biased and balanced cases, both assuming that $k = 1$ and $p = 1/2$. In both cases we notice that D converges to zero as fast as N^{-3}, and that the biased compensation method outperforms the balanced compensation method.

Example 10. Now consider Example 5, where the group is $\mathbb{Z}_N$ and $S = \{0, 1\}$. As we already noticed, for this example the biased and the balanced methods coincide. From the computations of the matrix A and of the vector B we obtain

$$\frac{B_h}{1 - A_{hh}} = \frac{2p(1-p)k^2\left(1 - \cos\left(\frac{2\pi}{N}i\right)\right)}{2pk(1-pk)\left(1 - \cos\left(\frac{2\pi}{N}i\right)\right)} = \frac{(1-p)k}{1-pk}\,.$$

Hence,

$$D = \frac{N^{-1}}{1 - \frac{(1-p)k}{1-pk}}\frac{(1-p)k}{1-pk} = N^{-2}\frac{k(1-p)}{1-k+N^{-1}(1-p)k}\,.$$

Also in this case, for $p \to 1$, or for $N \to +\infty$, we have that $D \to 0$. Notice that the speed of convergence to 0 with respect to N is lower than in the complete case.

In order to have a clearer insight into the behavior of D we make some further estimations by analyzing the two cases separately.

Let us start with the biased case. Notice that we have

$$\frac{B_h}{1 - A_{hh}} = \frac{B(\chi_h)}{1 - A(\chi_h)} = \frac{2p(1-p)[\hat{\pi}_{\tilde{P}}(\chi_0) - \Re[\hat{\pi}_{\tilde{P}}(\chi_h)]]}{1 - |1 - p + p\hat{\pi}_P(\chi_h)|^2}\,.$$

Assume that $\pi_P(j) = k_j$, and notice that we have

$$\hat{\pi}_P(\chi_h) = \sum_j k_j\chi_h(-j)\,,\ \hat{\pi}_{\tilde{P}}(\chi_h) = \sum_j k_j^2\chi_h(-j)\,.$$

Hence,

$$\hat{\pi}_{\tilde{P}}(\chi_0) - \Re[\hat{\pi}_{\tilde{P}}(\chi_h)] = \sum_j k_j(1 - \Re[\chi_h(-j)])\,,$$

while

$$\begin{aligned} 1 - |1 - p + p\hat{\pi}_P(\chi_h)|^2 &= -p^2 + 2p - p^2|\hat{\pi}_P(\chi_h)|^2 - 2p(1-p)\sum_j k_j \Re\chi_h(j) \\ &= 2p(1-p)\sum_j k_j(1 - \Re[\chi_h(j)]) + p^2(1 - |\hat{\pi}_P(\chi_h)|^2)\,. \end{aligned}$$

We have thus obtained that

$$\frac{B_h}{1 - A_{hh}} = \frac{2(1-p)\sum_j k_j^2(1 - \Re[\chi_h(-j)])}{2(1-p)\sum_j k_j(1 - \Re[\chi_h(-j)]) + p(1 - |\hat{\pi}_P(\chi_h)|^2)}\,. \tag{36}$$

This explicit expression allows us to estimate D. We have the following result.

PROPOSITION 6.3. *Consider a Cayley matrix P and let $\pi_P(j)$ be its generator. Let $M = \max\{\pi_P(j)\,|\, j = 1, \dots, N-1\}$. Then*

$$D \le N^{-2}\frac{M}{1-M}\,.$$

Proof. From (36) we can argue that

$$\frac{B_h}{1 - A_{hh}} \le \frac{\hat{\pi}_{\tilde{P}}(\chi_0) - \Re[\hat{\pi}_{\tilde{P}}(\chi_h)]}{1 - \Re[\hat{\pi}_P(\chi_h)]}.$$

Assume that $\pi_P(j) = k_j$, and notice that we have

$$\hat{\pi}_P(\chi_h) = \sum_j k_j \chi_h(-j)\,,\ \hat{\pi}_{\tilde{P}}(\chi_h) = \sum_j k_j^2 \chi_h(-j)\,.$$

Hence,

$$\frac{B_h}{1 - A_{hh}} \le \frac{\sum_j k_j^2(1 - \Re[\chi_h(j)])}{\sum_j k_j(1 - \Re[\chi_h(j)])} \le M\,.$$

The thesis now simply follows from Proposition 6.2. ◻

The key point of Proposition 6.3 is that if we have a sequence of consensus strategies indexed by N, for which M is bounded away from 1, then D will converge to 0 at least as fast as N^{-2}. Notice that this is in agreement with the two examples considered above.

We now proceed to analyze the balanced case. We can prove the following result.

PROPOSITION 6.4. *Denote*

$$M = \pi_{\bar{P}(I)}(0);$$

then

$$D \le N^{-2}\frac{1}{1-M}\,. \tag{37}$$

Proof. In the balanced case A_{hh} and B_h are defined in (26) and (27). We obtain

$$\frac{B_h}{1 - A_{hh}} = \frac{\hat{\pi}_{\bar{\rho}(I)}(\chi_h) - \left|\hat{\pi}_{\bar{\beta}}(\chi_h)\right|^2}{1 - \left|\hat{\pi}_{\bar{\beta}}(\chi_h)\right|^2}. \tag{38}$$

It follows from (38), using the inequality $0 \leq \hat{\pi}_{\bar{\rho}(I)}(\chi_h) \leq 1$, that

$$\frac{B_h}{1 - A_{hh}} \leq \hat{\pi}_{\bar{\rho}(I)}(\chi_h) \leq 1\,. \tag{39}$$

Notice now that

$$N^{-1} \sum_{j=1}^{N-1} \frac{B_j}{1 - A_{jj}} \leq N^{-1} \sum_{j=1}^{N-1} \hat{\pi}_{\bar{\rho}(I)}(\chi_j) \leq N^{-1} \sum_{j=0}^{N-1} \hat{\pi}_{\bar{\rho}(I)}(\chi_j) = \pi_{\bar{\rho}(I)}(0)\,. \tag{40}$$

The thesis now follows from Proposition 6.2 and estimations (39) and (40). □

Notice that M is strictly smaller than 1. It clearly depends on the matrix P but also, as opposed to the biased case, on the probability p. Let us analyze a simple case in more detail. Assume that the matrix P is defined as in (28). In this case, using (30), we have that

$$\begin{aligned} M &= \pi_{\bar{\rho}(I)}(0) = N^{-1}\text{trace } \bar{\rho}(I) \\ &= N^{-1}\text{trace } (\rho H^* H + (\xi - \rho)\text{diag } (H^* H) + \gamma I + \delta(H + H^*)) \\ &= (n-1)\xi + \gamma. \end{aligned}$$

Therefore M depends on n and p and k, but it does not depend on N. It thus follows that δ converges to 0 as fast as N^{-2} in the biased case.

7. Conclusions. In this paper we proposed some tools which allow us to evaluate the performance degradation due to failing transmission links in the average consensus algorithm. Though the tools proposed here seem to be very effective for the evaluation of the effect of packet drop in the data transmission between the agents in a consensus seeking problem, many problems are still to be investigated, such as the following:

1. The analysis of convergence has been carried out in a mean square sense. Concentration results can be obtained in certain cases (see [6]) and it would be important to study them in the context of packet drop models.
2. Many problems are still open in the general (non-Cayley) case, such as the evaluation of the mean distance of the limit from the average as a function of the number N of agents.
3. The analysis is still quite intricate and it is difficult to use in design. We expect that some interesting simplifications could occur when N tends to infinity. It is important to determine whether this is really the case and to exploit these simplifications in the design process.
4. The average consensus algorithm we considered is somehow memoryless. We expect that algorithms with memory in principle could yield better performance (consider for instance an algorithm which, when data is lost, can substitute it with its past version). It is important to understand whether adding memory will improve the performance or not.

REFERENCES

[1] L. Babai, *Spectra of Cayley graphs*, J. Combin. Theory Ser. B, 27 (1979), pp. 180–189.

[2] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, *Randomized gossip algorithms*, IEEE Trans. Inform. Theory, 52 (2006), pp. 2508–2530.

[3] R. Carli, F. Fagnani, A. Speranzon, and S. Zampieri, *Communication constraints in the average consensus problem*, Automatica, 44 (2008), pp. 671–684.

[4] R. Cogburn, *On products of random stochastic matrices*, in Random Matrices and Their Applications (Brunswick, Maine, 1984), Contemp. Math. 50, AMS, Providence, RI, 1986, pp. 199–213.

[5] R. Diestel, *Graph Theory*, Grad. Texts in Math. 173, Springer-Verlag, Heidelberg, 2005.

[6] F. Fagnani and S. Zampieri, *Randomized consensus algorithms over large scale networks*, IEEE J. Sel. Areas Commun., 26 (2008), pp. 634–649.

[7] F. R. Gantmacher, *The Theory of Matrices*, Chelsea, New York, 1959.

[8] C. Godsil and G. Royle, *Algebraic Graph Theory*, Grad. Texts in Math. 207, Springer-Verlag, New York, 2001.

[9] Y. Hatano and M. Mesbahi, *Agreement over random networks*, IEEE Trans. Automat. Control, 50 (2005), pp. 1867–1872.

[10] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1994.

[11] A. Jadbabaie, J. Lin, and A. S. Morse, *Coordination of groups of mobile autonomous agents using nearest neighbor rules*, IEEE Trans. Automat. Control, 48 (2003), pp. 988–1001.

[12] Z. Lin, B. Francis, and M. Maggiore, *State agreement for continuous-time coupled nonlinear systems*, SIAM J. Control Optim., 46 (2007), pp. 288–307.

[13] L. Moreau, *Stability of multiagent systems with time-dependent communication links*, IEEE Trans. Automat. Control, 50 (2005), pp. 169–182.

[14] R. Olfati-Saber, *Distributed Kalman filter with embedded consensus filters*, in Proceedings of the 44th IEEE Conference on Decision and Control and European Control Conference, 2005, pp. 8179–8184.

[15] R. Olfati-Saber and R. M. Murray, *Consensus problems in networks of agents with switching topology and time-delays*, IEEE Trans. Automat. Control, 49 (2004), pp. 1520–1533.

[16] R. Olfati-Saber, J. A. Fax, and R. M. Murray, *Consensus and cooperation in networked multi-agent systems*, Proceedings of the IEEE, 95 (2007), pp. 215–233.

[17] S. Patterson, B. Bamieh, and A. El Abbadi, *Distributed average consensus with stochastic communication failures*, in Proceedings of the 46th IEEE Conference on Decision and Control, 2007, pp. 4215–4220.

[18] W. Ren and R. W. Beard, *Consensus seeking in multiagent systems under dynamically changing interaction topologies*, IEEE Trans. Automat. Control, 50 (2005), pp. 655–661.

[19] A. Terras, *Fourier analysis on finite groups and applications*, London Math. Soc. Stud. Texts 43, Cambridge University Press, Cambridge, UK, 1999.

[20] J. Tsitsiklis, *Problems in Decentralized Decision Making and Computation*, Ph.D. thesis, Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, 1984.

SIAM J. Control Optim.
Vol. 48, No. 1, pp. 134–161

© 2009 Society for Industrial and Applied Mathematics

COORDINATION AND CONSENSUS OF NETWORKED AGENTS WITH NOISY MEASUREMENTS: STOCHASTIC ALGORITHMS AND ASYMPTOTIC BEHAVIOR*

MINYI HUANG† AND JONATHAN H. MANTON‡

Abstract. This paper considers the coordination and consensus of networked agents where each agent has noisy measurements of its neighbors' states. For consensus seeking, we propose stochastic approximation-type algorithms with a decreasing step size, and introduce the notions of mean square and strong consensus. Although the decreasing step size reduces the detrimental effect of the noise, it also reduces the ability of the algorithm to drive the individual states towards each other. The key technique is to ensure a trade-off for the decreasing rate of the step size. By following this strategy, we first develop a stochastic double array analysis in a two-agent model, which leads to both mean square and strong consensus, and extend the analysis to a class of well-studied symmetric models. Subsequently, we consider a general network topology, and introduce stochastic Lyapunov functions together with the so-called direction of invariance to establish mean square consensus. Finally, we apply the stochastic Lyapunov analysis to a leader following scenario.

Key words. multiagent systems, graphs, consensus problems, measurement noise, stochastic approximation, mean square convergence, almost sure convergence

AMS subject classifications. 93E03, 93E15, 94C15, 68R10

DOI. 10.1137/06067359X

1. Introduction.

The recent years have witnessed an enormous growth of research on the coordination and control of distributed multiagent systems, and specific topics appear in different forms such as swarming of honeybees, flocking of birds, migration of animals, synchronization of coupled oscillators, and formation of autonomous vehicles; see [48, 14, 17, 29, 43, 33] and the references therein. A common feature of these systems, which take diverse forms, is that the constituent agents need to maintain a certain coordination so as to cooperatively achieve a group objective, wherein the decision of individual agents is made with various constraints due to the distributed nature of the underlying system. The study of these multiagent models is crucial for understanding many complex phenomena related to animal behavior, and for designing distributed control systems.

For multiagent coordination, it is usually important to propagate shared information within the system by communication rules which may be supported by the specific interconnection structure between the agents. This is particularly important in cooperative control systems since they often operate in a dynamic environment, and the involved agents need to collectively acquire key information at the overall system level [38, 3]. In this context, of fundamental importance is the so-called consensus or agreement problem, where consensus means a condition where all the agents individually adjust their own value for an underlying quantity (e.g., a location as the destination of a robot team) so as to converge to a common value. For many prac-

*Received by the editors October 28, 2006; accepted for publication (in revised form) July 30, 2008; published electronically February 11, 2009. This work was partially supported by the Australian Research Council and the Australian National University.

http://www.siam.org/journals/sicon/48-1/67359.html

†Corresponding author. School of Mathematics and Statistics, Carleton University, Ottawa, ON K1S 5B6, Canada (mhuang@math.carleton.ca).

‡Department of Electrical and Electronic Engineering, The University of Melbourne, Parkville, VIC, 3010, Australia (j.manton@ieee.org).

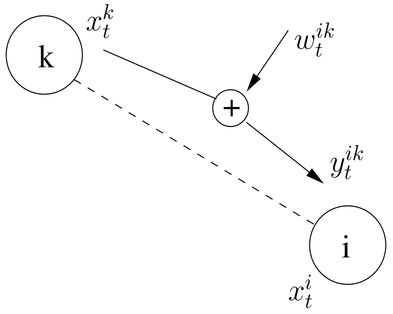


FIG. 1. *Measurement with additive noise w_t^{ik}.*

tical situations, the chief objective is to agree on the same state; the actual state is of secondary importance. In view of primarily being required to converge, one might suggest to simply set the agents' states to any fixed state. In reality, however, such a consensus protocol is trivial and less interesting; its more serious limitation is that this protocol is overly sensitive to small relative errors when the individual states initially have been very close to each other. For these reasons, in the literature, almost all consensus algorithms are constructed based on averaging rules, and this leads to good dynamic properties (such as good transient behavior and convergence) [23, 50, 6]. We mention that there has been a long history of research on consensus problems due to the broad connections of this subject with a wide range of disciplines including statistical decision theory, management science, distributed computing, ad hoc networks, biology [49, 20, 10, 18, 28, 26, 48], and the quickly developing area of multiagent control systems [3, 14, 17, 29, 33, 34, 43]. A comprehensive survey on consensus problems in multiagent coordination can be found in [38].

In the context of coordinating spatially distributed agents, a basic consensus model consists of a time-invariant network in which each agent updates its state by forming a convex combination of the states of its neighbors and itself [23, 6, 50], such that the iterates of all individual states converge to a common value. Starting from this formulation, many generalizations are possible. A variety of consensus algorithms has been developed to deal with delayed measurements [5, 31, 34], dynamic topologies [34], or unreliable (on/off) communication links (see the survey [38]). For convergence analysis, stochastic matrix analysis is an important tool [23], and in models with time-dependent communications, set-valued Lyapunov theory is useful [32].

In this paper, we are interested in consensus seeking in an uncertain environment where each agent can obtain only noisy measurements of the states of its neighbors; see Figure 1 for illustration. Such modeling reflects many practical properties in distributed networks. For instance, the interagent information exchange may involve the use of sensors, quantization [36, 37], and wireless fading channels, which makes it unlikely to have exact state exchange. We note that most previous research has used noise-free state iteration by assuming exact data exchange between the agents, with only a few exceptions (see, e.g., [51, 39, 9]). A least mean square optimization method was used in [51] to choose the constant coefficients in the averaging rule with additive noises so that the long term consensus error is minimized. In a continuous time consensus model [15], deterministic disturbances were included in the dynamics. In [9], multiplicative noises were introduced to model logarithmic quantization error. In [21, 42], convergence results were obtained for random graph based consensus problems, and [21] used an approach of stochastic stability. In earlier work [7, 46, 47], convergence of consensus problems was studied in a stochastic setting, but the interagent exchange of random messages was assumed to be error-free. In particular, Tsitsiklis, Bertsekas, and Athans [47] obtained consensus results via asynchronous stochastic gradient based algorithms for a group of agents minimizing their common cost function.

In models with noisy measurements, one may still construct an averaging rule with a constant coefficient matrix. However, the resulting evolution of the state vector dramatically differs from the noise-free case, leading to divergence. The reason is that the noise causes a steady drift of the agents' states during the iterates, which in turn prevents generating a stable group behavior.

To deal with the measurement noise, we propose a stochastic approximation-type algorithm with the key feature of a decreasing step size. The algorithm has a gradient descent interpretation. Our formulation differs from [47] since in the averaging rule of the latter, the exogenous term, which may be interpreted as a local noisy gradient of the agents' common cost, is assigned a controlled step size while the weights for the exact messages received from other agents are maintained to be above a constant level; such a separability structure enables the authors in [47] to obtain consensus with a sufficiently small constant step size for the gradient term, or with only an upper bound for the deceasing rate of the step size. In contrast, in our model the signal received from other agents is corrupted by additive noise (see Figure 1), and consequently in selecting the step size, it is critical to maintain a trade-off in attenuating the noise to prevent long term fluctuations and meanwhile ensuring a suitable stabilizing capability of the recursion so as to drive the individual states towards each other. To achieve this objective, the step size must be decreased neither too slowly, nor too quickly. It turns out, for proving mean square consensus via stochastic Lyapunov functions, that we may simply use the standard step size condition in traditional stochastic approximation algorithms. But in the stochastic double array analysis, some mild lower and upper bound conditions will be imposed on the step size.

We begin by analyzing a two-agent model. As it turns out, this simple model provides a rich structure for developing convergence analysis and motivates the solution to more general models. In this setup, the key technique is the stochastic double array analysis [45, 12]. Next, we extend the analysis to a class of symmetric models. In fact, many symmetric models have arisen in practical applications including platoons of vehicles, robot teams, unicycle pursuit models [30, 29], cooperative sensor network deployment for tracking [1] or sampling [25], and consensus problems [9]. Subsequently, to deal with a general network topology, we develop a stochastic Lyapunov analysis, and convergence is established under a connectivity condition for the associated undirected graph.

The paper is organized as follows. In section 2 we formulate the consensus problem in the setting of directed graphs and propose the consensus algorithm. Section 3 establishes convergence results in a two-agent model, and the analysis is extended to models with circulant symmetry in section 4. We develop stochastic Lyapunov analysis in section 5 and apply it to leader following in section 6. Section 7 presents numerical simulations, and section 8 concludes the paper.

2. Formulation of the stochastic consensus problem. We begin by considering directed graphs for modeling the spatial distribution of n agents. A directed graph (or digraph) $G = (\mathcal{N}, \mathcal{E})$ consists of a set of nodes $\mathcal{N} = \{1, 2, \dots, n\}$ and a set of edges $\mathcal{E} \subset \mathcal{N} \times \mathcal{N}$. An edge in G is denoted as an ordered pair (i, j), where

$i \neq j$ (so there is no edge between a node and itself) and i, j are called the initial and terminal node, respectively. A path (from i_1 to i_l) in G consists of a sequence of nodes $i_1, i_2, \ldots, i_l$, $l \geq 2$, such that $(i_k, i_{k+1}) \in \mathcal{E}$ for all $1 \leq k \leq l-1$. The digraph G is said to be strongly connected if for any two distinct nodes i and j, there exist a path from i to j and also a path from j to i.

For convenience of exposition, we often refer to node i as agent A_i. The two names, agent and node, will be used interchangeably. Agent A_k (resp., node k) is a neighbor of A_i (resp., node i) if $(k, i) \in \mathcal{E}$, where $k \neq i$. Denote the neighbors of node i by $\mathcal{N}_i \subset \mathcal{N}$. Note that any undirected graph[1] can be converted into a directed graph simply by splitting each edge in the former into two edges, one in each direction.

For agent A_i, let $x_t^i \in \mathbb{R}$ be its state at time $t \in \mathbb{Z}^+ = \{0, 1, 2, \ldots\}$. Denote the state vector $x_t = [x_t^1, \ldots, x_t^n]^T$. For each $i \in \mathcal{N}$, agent A_i receives noisy measurements of the states of its neighbors. Denote the resulting measurement by A_i of A_k's state by

$$y_t^{ik} = x_t^k + w_t^{ik}, \qquad t \in \mathbb{Z}^+, \quad k \in \mathcal{N}_i, \tag{1}$$

where $w_t^{ik} \in \mathbb{R}$ is the additive noise; see Figure 1. The underlying probability space is denoted by $(\Omega, \mathcal{F}, P)$. We shall call y_t^{ik} the observation of the state of A_k obtained by A_i, and we assume each A_i knows its own state x_t^i exactly. The additive noise w_t^{ik} in (1) reflects unreliable information exchange during interagent sensing and communication; see, e.g., [39, 2, 41] for related modeling.

(A1) The noises $\{w_t^{ik}, t \in \mathbb{Z}^+, i \in \mathcal{N}, k \in \mathcal{N}_i\}$ are independent and identically distributed (i.i.d.) with respect to the indices i, k, t and each w_t^{ik} has zero mean and variance $Q \geq 0$. The noises are independent of the initial state vector x_0 and $E|x_0|^2 < \infty$.

Condition (A1) means that the noises are i.i.d. with respect to both space (associated with neighboring agents) and time. We will begin with our analysis based on the above assumption for simplicity.

The state of each agent is updated by

$$x_{t+1}^i = (1 - a_t) x_t^i + \frac{a_t}{|\mathcal{N}_i|} \sum_{k \in \mathcal{N}_i} y_t^{ik}, \qquad t \in \mathbb{Z}^+, \tag{2}$$

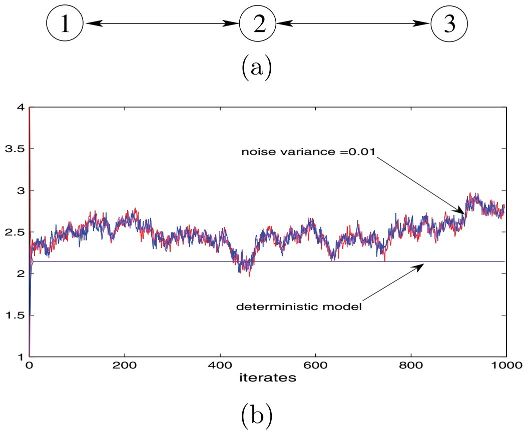
where $i \in \mathcal{N}$ and $a_t \in [0, 1]$ is the step size. This gives a weighted averaging rule in that the right-hand side is a convex combination of the agent's state and its $|\mathcal{N}_i|$ observations, where $|S|$ denotes the cardinality of a set S. The objective for the consensus problem is to select the sequence $\{a_t, t \geq 0\}$ so that the n individual states x_t^i, $i \in \mathcal{N}$, converge to a common limit in a certain sense.

To get some insight into algorithm (2), we rewrite it in the form

$$x_{t+1}^i = x_t^i + a_t (m_t^i - x_t^i), \tag{3}$$

where $m_t^i = (1/|\mathcal{N}_i|) \sum_{k \in \mathcal{N}_i} y_t^{ik}$. The structure of (3) is very similar to the recursion used in classical stochastic approximation algorithms in that $m_t^i - x_t^i$ provides a correction term controlled by the step size a_t. Indeed, by introducing a suitable local potential function, $m_t^i - x_t^i$ may be interpreted as the noisy measurement of a scaled negative gradient of the local potential along the direction x_t^i. A more detailed discussion will be presented in section 5 when developing the stochastic Lyapunov

[1] The edge in an undirected graph is denoted as an unordered pair.



(a)

(b)

Fig. 2. (a) *The three nodes.* (b) *In the noise-free case, the states of the nodes quickly converge to the same constant level* ≈ 2.143*. Under Gaussian measurement noises with variance* $\sigma^2 = 0.01$*, the three state trajectories have large fluctuations.*

analysis. Due to the noise contained in $\{m_t^i, t \geq 0\}$, each state x_t^i will fluctuate randomly. These fluctuations will not die off if a_t does not converge to 0. For illustration, we introduce an example as follows.

Example 1. *Consider a strongly connected digraph with* $\mathcal{N} = \{1, 2, 3\}$*, as in Figure* 2(a)*, where* $\mathcal{N}_1 = \{2\}$*,* $\mathcal{N}_2 = \{1, 3\}$*, and* $\mathcal{N}_3 = \{2\}$*. We follow the measurement model* (1)*, and the states are updated by* $x_{t+1}^1 = (x_t^1 + y_t^{12})/2$*,* $x_{t+1}^2 = (x_t^2 + y_t^{21} + y_t^{23})/3$*, and* $x_{t+1}^3 = (x_2^3 + y_t^{32})/2$*,* $t \geq 0$*. The i.i.d. Gaussian noises* w_t^{ik} *satisfy* (A1) *with variance* $\sigma^2 = 0.01$*.*

The simulation for Example 1 takes the initial condition $[x_0^1, x_0^2, x_0^3] = [4, 1, 2]$. For the noise-free case, we change the state update rule in Example 1 by replacing each y_t^{ik} by x_t^k, which results in a standard rule in the literature; see, e.g., [23]. Figure 2(b) shows that measurement noises cause a dramatic loss of convergence. In fact, by recasting to the form (2), the algorithm in Example 1 essentially takes the step size $a^{(i)} = |\mathcal{N}_i|/(|\mathcal{N}_i| + 1)$ for node i to give equal weights $1/(|\mathcal{N}_i| + 1)$ to $|\mathcal{N}_i| + 1$ nodes; for instance, we may rewrite $x_{t+1}^2 = (x_t^2 + y_t^{21} + y_t^{23})/3$ as $x_{t+1}^2 = x_t^2 + a^{(2)}[(y_t^{21} + y_t^{23})/2 - x_t^2]$, where $a^{(2)} = 2/3$.

With the aim of obtaining a stable behavior for the agents, we make the following assumption.

(A2) (i) The sequence $\{a_t, t \geq 0\}$ satisfies $a_t \in [0, 1]$, and (ii) there exists $T_0 \geq 1$ such that

$$\alpha t^{-\gamma} \leq a_t \leq \beta t^{-\gamma} \tag{4}$$

for all $t \geq T_0$, where $\gamma \in (0.5, 1]$ and $0 < \alpha \leq \beta < \infty$.

By requiring $a_t > \alpha t^{-\gamma}$ for $t \geq T_0$ with a suitable T_0, we may take large values for α while still ensuring $a_t \in [0, 1]$, $t \geq T_0$. This offers more flexibility in selecting the step size sequence. Here $\{a_t, t < T_0\}$ may be chosen freely as long as $a_t \in [0, 1]$; this resulting algorithm gives a convex combination at all times in the averaging rule as in conventional consensus algorithms. The parameters $T_0, \alpha, \beta, \gamma$ will be treated as fixed constants associated with $\{a_t, t \geq 0\}$. Note that (A2) implies the following weaker condition.

(A2′) (i) The sequence $\{a_t, t \geq 0\}$ satisfies $a_t \in [0,1]$, and (ii) $\sum_{t=0}^{\infty} a_t = \infty$, $\sum_{t=0}^{\infty} a_t^2 < \infty$.

Notice that (A2′)(ii) is a typical condition used in classical stochastic approximation theory [11, 24]. In the subsequent sections, the double array analysis will be developed based on the slightly stronger assumption (A2) while (A2′) will be used for the stochastic Lyapunov analysis.

The vanishing rate of $\{a_t, t \geq 0\}$ is crucial for consensus. When $a_t \to 0$ in (2), the signal x_t^k (contained in y_t^{ik}), as the state of A_k, is attenuated together with the noise. Hence, a_t cannot decrease too fast since, otherwise, the agents may prematurely converge to different individual limits.

Since the averaging rule (2) can be considered a stochastic approximation algorithm [27, 4], we may apply the standard method of analysis to it; namely, we can average out the noise component in (2) to derive an associated ordinary differential equation (ODE) system

$$\frac{dx^i(s)}{ds} = (1/|\mathcal{N}_i|) \sum_{k \in \mathcal{N}_i} x^k(s) - x^i(s), \quad s \geq 0, \quad i \in \mathcal{N}. \tag{5}$$

The important feature of the ODE system (5) is that it has an equilibrium set as a linear subspace of $\mathbb{R}^n$, instead of a singleton. This indicates more uncertain asymptotic behavior in the state evolution of the stochastic consensus algorithm due to the lack of a *single* equilibrium point generating the attracting effect, and is in contrast to typical stochastic approximation algorithms where the associated ODE usually has a single equilibrium, at least locally.

We introduce some definitions to characterize the asymptotic behavior of the agents.

Definition 2 (weak consensus). *The agents are said to reach weak consensus if* $E|x_t^i|^2 < \infty$, $t \geq 0$, $i \in \mathcal{N}$, *and* $\lim_{t\to\infty} E|x_t^i - x_t^j|^2 = 0$ *for all distinct* $i, j \in \mathcal{N}$.

Definition 3 (mean square consensus). *The agents are said to reach mean square consensus if* $E|x_t^i|^2 < \infty$, $t \geq 0$, $i \in \mathcal{N}$, *and there exists a random variable* x^* *such that* $\lim_{t\to\infty} E|x_t^i - x^*|^2 = 0$ *for all* $i \in \mathcal{N}$.

Definition 4 (strong consensus). *The agents are said to reach strong consensus if there exists a random variable* x^* *such that with probability* 1 *(w.p.*1*) and for all* $i \in \mathcal{N}$, $\lim_{t\to\infty} x_t^i = x^*$.

It is obvious that mean square consensus implies weak consensus. If a sequence converges w.p.1, we also say it converges almost surely (a.s.). Note that for both mean square and strong consensus, the states x_t^i, $i \in \mathcal{N}$, must converge to a common limit, which may depend on the initial states, the noise sequence, and the consensus algorithm itself.

2.1. The generalization to vector states.

It is straightforward to generalize the results of this paper to the case of vector individual states $\mathbf{x}_t^k \in \mathbb{R}^d$, where $d > 1$, and (1)–(2) may be extended to the vector case by taking a vector noise term. For the vector version of (2), we see that the d components in $\mathbf{x}_t^k$ are decoupled during iteration and may be treated separately. Throughout this paper, we consider only scalar individual states.

3. Convergence in a two-agent model.

We begin by analyzing a two-agent model, which will provide interesting insight into understanding consensus seeking in a noisy environment. The techniques developed for such a system will provide motivation for analyzing more general models. The rich structure associated with this seemingly simple model well justifies a careful investigation.

3.1. Mean square consensus. Let (1)–(2) be applied by the two agents where $\mathcal{N} = \{1, 2\}$. In the subsequent analysis, a key step is to examine the evolution of the difference $\xi_t = x_t^1 - x_t^2$ between the two states. We notice the relation

$$\xi_{t+1} = (1 - 2a_t)\xi_t + a_t v_t, \qquad t \geq 0, \tag{6}$$

where $v_t = w_t^{12} - w_t^{21}$. By inequality (4), we may find an integer $T_1 > T_0$ such that

$$1 - 2\alpha t^{-\gamma} \geq 1 - 2a_t > 0 \qquad \text{for all} \quad t \geq T_1. \tag{7}$$

In the estimate below, we start with T_1 as the initial time. It follows from (6) that

$$\xi_{t+1} = \prod_{i=T_1}^{t} (1 - \bar{a}_i)\xi_{T_1} + \sum_{k=T_1}^{t} \left[\prod_{i=k+1}^{t} (1 - \bar{a}_i) \right] a_k v_k, \qquad t \geq T_1, \tag{8}$$

where $\bar{a}_t = 2a_t$. Define

$$\Pi_{l,k} = \prod_{i=k+1}^{l} (1 - \bar{a}_i) a_k, \tag{9}$$

where $l > k \geq T_1$. By convention, $\Pi_{k,k} = a_k$.

Lemma 5. *Let $\Pi_{l,k}$ be defined by* (9) *with $k \leq l$ and assume* (A2).

(i) *If $\gamma = 1$, we have*

$$\Pi_{l,k} \leq \exp\left\{ -2\alpha \sum_{t=k+1}^{l} t^{-1} \right\} \frac{\beta}{k} \leq \frac{\beta(k+1)^{2\alpha}}{k(l+1)^{2\alpha}}. \tag{10}$$

(ii) *If $1/2 < \gamma < 1$, we have*

$$\Pi_{l,k} \leq \exp\left\{ \frac{-2\alpha}{1-\gamma}[(l+1)^{1-\gamma} - (k+1)^{1-\gamma}] \right\} \frac{\beta}{k^\gamma}. \tag{11}$$

Proof. First, for the case $k < l$, it is obvious that

$$\Pi_{l,k} \leq \left(1 - \frac{2\alpha}{l^\gamma}\right) \cdots \left(1 - \frac{2\alpha}{(k+1)^\gamma}\right) \frac{\beta}{k^\gamma}. \tag{12}$$

By the fact $\ln(1 - x) < -x$ for all $x \in (0, 1)$, it follows that

$$\left(1 - \frac{2\alpha}{l^\gamma}\right) \cdots \left(1 - \frac{2\alpha}{(k+1)^\gamma}\right) \leq \exp\left\{ -2\alpha \sum_{t=k+1}^{l} t^{-\gamma} \right\}. \tag{13}$$

By (12)–(13), we get (i) and (ii) for $k < l$. Clearly, (i) and (ii) hold for $k = l$. □

Let $\{c(t), t \geq t_0\}$ and $\{h(t), t \geq t_0\}$ be two sequences of real numbers indexed by integers $t \geq t_0$, and $h(t) > 0$ for all $t \geq t_0$. Denote $c(t) = O(h(t))$ (resp., $c(t) = o(h(t))$) if $\overline{\lim}_{t\to\infty} |c(t)|/h(t) \leq C_d < \infty$ (resp., $\lim_{t\to\infty} |c(t)|/h(t) = 0$). Here C_d is called a dominance constant in the relation $c(t) = O(h(t))$. In practice, it is desirable to take a value for C_d as small as possible.

LEMMA 6. *Under* (A2)*, we have the upper bound estimate* (i) *if* $\gamma = 1$,

$$\sum_{k=T_1}^{t} \Pi_{t,k}^2 = \begin{cases} O(t^{-4\alpha}) & \textit{if} \quad 0 < \alpha < 1/4, \\ O(t^{-1}\ln t) & \textit{if} \quad \alpha = 1/4, \\ O(t^{-1}) & \textit{if} \quad \alpha > 1/4, \end{cases} \tag{14}$$

where T_1 *is specified in* (7)*, and* (ii) *if* $1/2 < \gamma < 1$,

$$\sum_{k=T_1}^{t} \Pi_{t,k}^2 = O(t^{-\gamma}). \tag{15}$$

Proof. See the appendix. □

Remark. We give some discussions on estimating the dominance constant C_d for Lemma 6. For (14), when $\alpha \neq 1/4$ but is close to $1/4$ from the left (resp., right), our estimation method shows that we need to take a large C_d associated with $O(t^{-4\alpha})$ (resp., t^{-1}). For the case $\alpha = 1/4$ in (14), we may take $C_d = \beta^2$. For (15), we take $C_d = 4\alpha$, regardless of the value of $\gamma \in (1/2, 1)$.

COROLLARY 7. *Let* $\{\tilde{a}_t, t \geq 1\}$ *be a sequence such that* (i) $\tilde{a}_t \in [0,1]$ *and* (ii) *there exists* $\gamma_0 \in (0, 1/2)$ *such that* $\tilde{\alpha} t^{-\gamma_0} \leq \tilde{a}_t \leq \tilde{\beta} t^{-\gamma_0}$*, where* $\tilde{\alpha} > 0$*. Denote* $\tilde{\Pi}_{l,k} = \prod_{i=k+1}^{l}(1-\tilde{a}_i)\tilde{a}_k$*,* $l \geq k \geq 1$*. Then for any fixed* $\tilde{T}_1 \geq 1$*,* $\sum_{k=\tilde{T}_1}^{t} \tilde{\Pi}_{t,k}^2 = O(t^{-\gamma_0})$.

Proof. First, (11) is still valid after replacing γ (resp., $\Pi_{l,k}$) by γ_0 (resp., $\tilde{\Pi}_{l,k}$). The argument in proving (15) can be repeated when γ is replaced by γ_0, which leads to the corollary. □

THEOREM 8. *Suppose* (A1)–(A2) *hold for the system of two agents, and* x_t^1*,* x_t^2 *are updated according to algorithm* (2)*. Then there exists a random variable* x^* *such that* $\lim_{t\to\infty} E|x_t^i - x^*|^2 = 0$ *for* $i = 1, 2$*, which implies mean square consensus.*

Proof. First, denote $z_t = (x_t^1 + x_t^2)/2$ and $\tilde{w}_t = (w_t^{12} + w_t^{21})/2$ for $t \geq 0$. It is easy to check that

$$z_{t+1} = z_t + a_t \tilde{w}_t, \qquad t \geq 0, \tag{16}$$

which leads to $z_{t+1} = z_0 + \sum_{k=0}^{t} a_k \tilde{w}_k$. Since $\sum_{t=0}^{\infty} a_t^2 < \infty$, there exists a random variable z^* such that $\lim_{t\to\infty} E|z_t - z^*|^2 = 0$.

Now we estimate $\xi_t = x_t^1 - x_t^2$. We see that

$$E\xi_{t+1}^2 \leq 2\left(E\xi_{T_1}^2 \prod_{k=T_1}^{t} |1 - 2a_k|^2 + \sup_{k \geq T_1} Ev_k^2 \times \sum_{k=T_1}^{t} \Pi_{t,k}^2 \right).$$

By Lemma 6, $\lim_{t\to\infty} E\xi_{t+1}^2 = 0$. Then mean square consensus follows easily. □

The i.i.d. noise assumption in Theorem 8 may be relaxed to independent noises with zero mean and uniformly bounded variances.

We use this two-agent model to illustrate the importance of a trade-off in the decreasing rate of a_t. To avoid triviality, assume the noise variance $Q > 0$ in (A1).

First, let $\gamma_0 \in (0, 1/2)$ and $a_0 = 0$, $a_t = t^{-\gamma_0}$ for $t \geq 1$, which decreases more slowly than in (4). By (16), it follows that $\lim_{t\to\infty} E|z_t|^2 = \infty$. Let ξ_t be given by (6). By Corollary 7, we can show $\lim_{t\to\infty} \xi_t^2 = 0$. So we conclude that this too-slowly-decreasing step size causes divergence of x_t^1 and x_t^2 due to inadequately attenuated noise, although they reach weak consensus since $\lim_{t\to\infty} \xi_t^2 = 0$.

Next, we take $\gamma_1 > 1$ and $a_0 = 0$, $a_t = t^{-\gamma_1}$ for $t \geq 1$, which decreases faster than in (4). Then there exists a random variable z^* such that $\lim_{t\to\infty} E|z_t - z^*|^2 = 0$.

Furthermore, by the fact $\prod_{i=2}^{\infty}(1-2\bar{a}_i)>0$, we obtain from (8) that there exists a random variable ξ^* such that $\lim_{t\to\infty} E|\xi_t-\xi^*|^2=0$ and $E|\xi^*|^2>0$. So x_t^1 and x_t^2 both converge in mean square. But the state gap ξ_t cannot be asymptotically eliminated due to the excessive loss of the stabilizing capability, associated with the homogenous part of (6), when a_t decreases too quickly.

3.2. Strong consensus. So far we have shown that the two states converge in mean square to the same limit. It is well known that in classical stochastic approximation theory [11, 24], similarly structured algorithms have sample path convergence properties under reasonable conditions. It is tempting to analyze sample path behavior in this consensus context. The analysis below moves towards this objective. The following lemma is instrumental.

LEMMA 9 (see [45]). *Let $\{w, w_t, t\ge 1\}$ be i.i.d. real-valued random variables with zero mean, and $\{a_{ki}, 1\le i\le l_k \uparrow \infty, k\ge 1\}$ a double array of constants. Assume* (i) $\max_{1\le i\le l_k}|a_{ki}| = O((l_k^{1/p}\log k)^{-1})$, $0<p\le 2$, *and* $\log l_k = o(\log^2 k)$, *and* (ii) $E|w|^p<\infty$. *Then* $\lim_{k\to\infty}\sum_{i=1}^{l_k} a_{ki}w_i = 0$ *a.s.*

This lemma is an immediate consequence of Theorem 4 and Corollary 3 in [45, pp. 331 and 340], which deal with the sum of random variables with weights in a double array.

Now we need to estimate the magnitude of the individual terms $\Pi_{t,k}$. Note that for each $t>T_1$, $\Pi_{t,k}$ is defined for k starting from T_1 up to t. Hereafter, for notational brevity, we make a convention about notation by setting $\Pi_{t,k}\equiv 0$ for $1\le k<T_1$ when $t\ge T_1$, and $\Pi_{t,k}\equiv 0$ for $1\le k\le t$ when $1\le t<T_1$. After this extension, all the entries $\Pi_{t,k}$ constitute a triangular array.

LEMMA 10. *For case* (i) *with* $\gamma=1$, *under* (A2) *we have*

$$\sup_{1\le k\le t}\Pi_{t,k} = \begin{cases} O(t^{-2\alpha}) & \text{if } \ 0<\alpha<1/2,\\ O(t^{-1}) & \text{if } \ \alpha\ge 1/2,\end{cases} \tag{17}$$

and for case (ii) *with* $1/2<\gamma<1$, *we have* $\sup_{1\le k\le t}\Pi_{t,k}=O(t^{-\gamma})$.

Proof. By use of (10), it is easy to obtain the bound for case (i). Now we give the proof for case (ii). By Lemma 5(ii), it follows that

$$\Pi_{t,k}\le e^{-\delta(t+1)^{1-\gamma}}e^{\delta(k+1)^{1-\gamma}}\frac{\beta}{k^\gamma}\le e^{-\delta(t+1)^{1-\gamma}}\max_{1\le k\le t}e^{\delta(k+1)^{1-\gamma}}\frac{\beta}{k^\gamma},$$

where $\delta=2\alpha/(1-\gamma)$. Denote the function $f(s)=e^{\delta(s+1)^{1-\gamma}}(\beta/s^\gamma)$, where the real number $s\in[1,\infty)$. By calculating the derivative $f'(s)$, it can be shown that for all $s\ge s_0=[1+\frac{\gamma}{\delta(1-\gamma)}]^{1/(1-\gamma)}$, $f'(s)>0$. Hence there exists $c_0>0$ independent of t such that

$$\max_{1\le k\le t}e^{\delta(k+1)^{1-\gamma}}\frac{\beta}{k^\gamma}\le\max_{s\in[1,t]}f(s)\le c_0\vee\left(e^{\delta(t+1)^{1-\gamma}}\frac{\beta}{t^\gamma}\right),$$

which implies that $\sup_{1\le k\le t}\Pi_{t,k}=O(t^{-\gamma})$. This completes the proof. □

THEOREM 11. *Assume all conditions in Theorem* 8 *hold and, in addition,* $\alpha>1/4$ *in the case* $\gamma=1$. *Then we have* (a) z_t *converges a.s.,* (b) $\lim_{t\to\infty}\xi_t=0$ *a.s., and* (c) *the two sequences* $\{x_t^1, t\ge 0\}$ *and* $\{x_t^2, t\ge 0\}$ *converge to the same limit a.s., which implies strong consensus.*

Proof. Recall that $z_{t+1}=z_0+\sum_{k=0}^t a_k\tilde{w}_k$ for $t\ge 0$, where $\tilde{w}_t=(w_t^{12}+w_t^{21})/2$. Since $\{\tilde{w}_k, k\ge 0\}$ is a sequence of independent random variables with zero mean

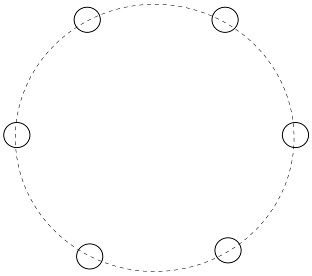


FIG. 3. *A symmetric ring network where each agent has two neighbors.*

and satisfies $\sum_{k=0}^{\infty} E|a_k\tilde{w}_k|^2 \leq \sup_k E|\tilde{w}_k|^2(\sum_{k=0}^{\infty} a_k^2) < \infty$, by the Khintchine–Kolmogorov convergence theorem [13], $\sum_{k=0}^{\infty} a_k\tilde{w}_k$ converges a.s. Hence assertion (a) holds.

Now we prove (b). By Lemma 10 we have

$$\sup_{1\leq k\leq t} \Pi_{t,k} = O((t^{1/2}\log t)^{-1}), \tag{18}$$

whenever $\alpha > 0$ (resp., $\alpha > 1/4$) in the case $1/2 < \gamma < 1$ (resp., $\gamma = 1$). To apply Lemma 9, we take $l_k = k$ and $p = 2$, which combined with (18) yields $\lim_{t\to\infty} \sum_{k=1}^{t} \Pi_{t,k}v_k = 0$, a.s. Hence $\lim_{t\to\infty} \xi_t = 0$ a.s. by (8), and (b) follows. Assertion (c) immediately follows from (a) and (b). □

The requirement $\alpha > 1/4$, associated with $\gamma = 1$, is a mild condition, and from an algorithmic point of view, it is not an essential restriction since in applications $\{a_t, t \geq 0\}$ is a sequence to be designed. In fact, by a slightly more complicated procedure, the restriction $\alpha > 1/4$ can be removed; see the more recent work [22].

4. Models with symmetric structures.

We continue to consider models where the neighboring relation for the n agents displays a certain symmetry. A simple example is shown in Figure 3 with ring-coupled agents each having two neighbors.

We specify the associated digraph as follows. First, the n nodes are listed by the order $1, 2, \ldots, n$. The ith node has a neighbor set $\mathcal{N}_i$ listed as $(\alpha_1^i, \ldots, \alpha_K^i)$ as a subset of $\{1, \ldots, n\}$. The constant $K \geq 1$ denotes the number of neighbors, which is the same for all agents. Then the $(i+1)$th node's neighbors are given by $(\alpha_1^i+1, \ldots, \alpha_K^i+1)$. In other words, by incrementing each α_k^i (associated with A_i) by one, where $1 \leq k \leq K$, we obtain the neighbor set for node $i+1$, and after a total of n steps, we retrieve node i and its neighbors $\mathcal{N}_i$. In fact, the underlying digraph may be realized by arranging the n nodes sequentially on a ring and adding the edges accordingly. For this reason, we term the fulfillment of the above incrementing rule as the circulant invariance property of the digraph. In this section, if an index (e.g., $\alpha_k^i + 1$) for a node or agent exceeds n, we identify it as an integer between 1 and n by taking $\mathrm{mod}(n)$.

Notice that the above symmetry assumption does not ensure the strong connectivity of the digraph. For illustration, consider a digraph with the set of nodes $\mathcal{N} = \mathcal{S}_1 \cup \mathcal{S}_2$, where $\mathcal{S}_1 = \{1, 3, 5\}$ and $\mathcal{S}_2 = \{2, 4, 6\}$. All nodes inside each $\mathcal{S}_i$, $i = 1, 2$, are neighbors to each other, but there exists no edge between two nodes with one in $\mathcal{S}_1$ and the other in $\mathcal{S}_2$. This digraph has the circulant invariance property without connectivity. Throughout this section, we make the following assumption.

(A3) The digraph $G = (\mathcal{N}, \mathcal{E})$ has the circulant invariance property and strong connectivity.

Define the centroid of the state configuration $(x_t^1, \ldots, x_t^n)$ as $z_t = (1/n)\sum_{i=1}^n x_t^i$. Under (A3), it is easy to show that z_t satisfies

$$z_{t+1} = z_t + (a_t/(nK)) \sum_{i\in\mathcal{N}} \sum_{k\in\mathcal{N}_i} w_t^{ik}, \qquad t \geq 0. \tag{19}$$

Lemma 12. *Under* (A1)–(A3)*, the sequence* $\{z_t, t \geq 0\}$ *converges in mean square and a.s.*

Proof. The lemma may be proved by the same method as in analyzing $\{z_t, t \geq 0\}$ in Theorems 8 and 11 for the two-agent model, and the details are omitted. □

We further denote the difference between x_t^{i+1} and x_t^i by

$$\xi_t^i = x_t^{i+1} - x_t^i, \qquad 1 \leq i \leq n. \tag{20}$$

Note that i and $i+1$ are two consecutively labelled agents, unnecessarily being neighbors to each other. By our convention, x_t^{n+1} is identified as x_t^1. Thus $\xi_t^n = x_t^1 - x_t^n$. The variables ξ_t^i, $1 \leq i \leq n$, are not linearly independent. Recall that $|\mathcal{N}_i| = K$ for all $i \in \mathcal{N}$. Specializing algorithm (2) to the model of this section, we have

$$x_{t+1}^i = (1-a_t)x_t^i + (a_t/K)\sum_{k\in\mathcal{N}_i}(x_t^k + w_t^{ik}) \tag{21}$$

for each $i \in \mathcal{N}$, and

$$\begin{aligned} x_{t+1}^{i+1} &= (1-a_t)x_t^{i+1} + (a_t/K)\sum_{k\in\mathcal{N}_{i+1}}(x_t^k + w_t^{i+1,k}) \\ &= (1-a_t)x_t^{i+1} + (a_t/K)\sum_{k\in\mathcal{N}_i}(x_t^{k+1} + w_t^{i+1,k+1}), \end{aligned} \tag{22}$$

where we obtain (22) by use of the circulant invariance of the neighboring relation.

By subtracting both sides of (22) by (21), we get the dynamics

$$\xi_{t+1}^i = (1-a_t)\xi_t^i + (a_t/K)\sum_{k\in\mathcal{N}_i}\xi_t^k + (a_t/K)\tilde{w}_t^i, \qquad i \in \mathcal{N}, \tag{23}$$

where

$$\tilde{w}_t^i = \sum_{k\in\mathcal{N}_i}\tilde{w}_t^{i,k}, \qquad \tilde{w}_t^{i,k} = w_t^{i+1,k+1} - w_t^{i,k} \tag{24}$$

with $k \in \mathcal{N}_i$ for $\tilde{w}_t^{i,k}$.

Lemma 13. *Let* ξ_t^i *and* $\tilde{w}_t^i$ *be defined by* (20) *and* (24)*, respectively. Under* (A3) *we have the zero-sum property:* $\sum_{i\in\mathcal{N}}\xi_t^i = 0$ *and* $\sum_{i\in\mathcal{N}}\tilde{w}_t^i = 0$ *for all* $t \geq 0$.

Proof. The first equality holds by the definition of ξ_t^i, $1 \leq i \leq n$. We now prove the second equality:

$$\begin{aligned} \sum_{i\in\mathcal{N}}\tilde{w}_t^i &= \sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{N}_i} w_t^{i+1,k+1} - \sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{N}_i} w_t^{i,k} \\ &= \sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{N}_i} w_t^{i,k} - \sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{N}_i} w_t^{i,k} \\ &= 0, \end{aligned} \tag{25}$$

where we get (25) by the circulant invariance property. □

Before further analysis, we introduce the $n \times n$ stochastic matrix

$$M(a) = I + aM^c, \qquad a \in [0,1]. \tag{26}$$

The circulant matrix M^c is given in the form

$$M^c = \begin{bmatrix} -1 & c_1 & c_2 & \cdots & c_{n-1} \\ c_{n-1} & -1 & c_1 & \ddots & \vdots \\ c_{n-2} & c_{n-1} & -1 & \ddots & c_2 \\ \vdots & \ddots & \ddots & \ddots & c_1 \\ c_1 & \cdots & c_{n-2} & c_{n-1} & -1 \end{bmatrix},$$

where $M^c_{ii} = -1$ for $1 \le i \le n$, and for $2 \le k \le n$,

$$M^c_{1k} = c_{k-1} = \begin{cases} 1/K & \text{if } k \in \mathcal{N}_1, \\ 0 & \text{otherwise.} \end{cases}$$

Since M^c is a circulant matrix [16], it is well defined after the first row is determined. In fact, both M^c and $M(a)$ are circulant matrices.

PROPOSITION 14. *Under* (A3), *$M(a)$ is doubly stochastic for any $a \in [0,1]$; i.e., both $M(a)$ and $[M(a)]^T$ are stochastic matrices, and $M(a)$ is irreducible for $a \in (0,1]$.*

Proof. All row and column sums in $M(a)$ are equal to one. Hence $M(a)$ is doubly stochastic. Since G is strongly connected, $M(a)$ is irreducible for $a > 0$. □

Define $\xi_t = [\xi^1_t, \dots, \xi^n_t]^T$ and $\tilde{w}_t = [\tilde{w}^1_t, \dots, \tilde{w}^n_t]^T$. We can check that ξ_t satisfies

$$\xi_{t+1} = M(a_t)\xi_t + (a_t/K)\tilde{w}_t, \quad t \ge 0. \tag{27}$$

The following lemma plays an essential role for the stability analysis of (27).

LEMMA 15. *Assume* (A2)–(A3) *hold, and the real vector $\theta = [\theta_1, \dots, \theta_n]^T$ has a zero column sum, i.e., $\sum_{i=1}^n \theta_i = 0$. Then for all $t \ge k \ge 0$, we have*

(i) *The column sum of $M(a_t)\dots M(a_k)\theta$ is zero, i.e., $\sum_{i=1}^n M^\theta_{t,k}(i) = 0$, where we denote $M^\theta_{t,k} = [M^\theta_{t,k}(1), \dots, M^\theta_{t,k}(n)]^T = M(a_t)\dots M(a_k)\theta$.*

(ii) *There exist constants $\delta^* \in (0,1)$ and $T_2 > 0$, both independent of θ, such that*

$$|M(a_t)\dots M(a_k)\theta| \le |(1-\delta^* a_t)\dots(1-\delta^* a_k)\theta|$$

for all $t \ge k \ge T_2$, where T_2 is chosen such that $a_t \le 1/2$ for all $t \ge T_2$.

Proof. The matrix $M(a_k)$, $k \ge 0$, is doubly stochastic by Proposition 14. Then θ having a zero column sum implies $M(a_k)\theta$ has a zero column sum. Repeating this argument, we obtain part (i).

We now prove (ii). First, let $\omega_n = e^{2\pi \mathbf{i}/n}$, where $\mathbf{i} = \sqrt{-1}$ is the imaginary unit, and denote

$$F_n = \frac{1}{\sqrt{n}} \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega_n & \omega_n^2 & \cdots & \omega_n^{n-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \omega_n^{n-1} & \omega_n^{2(n-1)} & \cdots & \omega_n^{(n-1)(n-1)} \end{pmatrix},$$

which is the so-called Fourier matrix of order n and satisfies $F_n^* F_n = I$, where F_n^* is the conjugate transpose of F_n. For $a \in [0,1]$, we introduce the polynomial

$$\varphi(a,z) = (1-a) + a(c_1 z + c_2 z^2 + \cdots + c_{n-1} z^n).$$

By well-known results for circulant matrices [16, 8], the n eigenvalues $\{\lambda_{1,t}, \ldots, \lambda_{n,t}\}$ of $M(a_t)$ are given by $\lambda_{k,t} = \varphi(a_t, \omega_n^{k-1})$ for $1 \le k \le n$. Obviously, $\lambda_{1,t} = 1$. Furthermore, $M(a_t)$ may be diagonalized in the form $M(a_t) = F_n^* \times \operatorname{diag}(\lambda_{1,t}, \ldots, \lambda_{n,t}) \times F_n$.

It is easy to verify that

$$\begin{aligned} M(a_t) \ldots M(a_k) &= F_n^* \times \Pi_{j=k}^t \operatorname{diag}(\lambda_{1,j}, \ldots, \lambda_{n,j}) \times F_n \\ &= F_n^* \times \Pi_{j=k}^t \operatorname{diag}(0, \lambda_{2,j}, \ldots, \lambda_{n,j}) \times F_n + (1/n) 1_n 1_n^T. \end{aligned}$$

Since $1_n 1_n^T \theta = 0$ for any θ with a zero column sum, we have

$$M(a_t) \ldots M(a_k)\theta = F_n^* \times \Pi_{j=k}^t \operatorname{diag}(0, \lambda_{2,j}, \ldots, \lambda_{n,j}) \times F_n \theta. \tag{28}$$

Notice that we may write $\varphi(a, w_n^{k-1}) = 1 + c_{k,1} a + \mathbf{i} c_{k,2} a$ for $2 \le k \le n$, where $c_{k,1}$ and $c_{k,2}$ are constants independent of a. For $0 < a < 1$, the matrix $M(a)$ is irreducible and aperiodic,[2] and hence for $2 \le k \le n$, $|\varphi(a, w_n^{k-1})| < \lambda_{1,t} = 1$; the reader is referred to [40] for additional details on spectral theory of stochastic matrices. Then we necessarily have $c_{k,1} < 0$, and in addition, for $0 < a < 1$,

$$|\varphi(a, \omega_n^{k-1})|^2 = (1 + c_{k,1} a)^2 + c_{k,2}^2 a^2 < 1, \quad 2 \le k \le n. \tag{29}$$

By taking $a \uparrow 1$ in (29), we get $-2 \le c_{k,1} < 0$, $|c_{k,2}| \le 1$, and $c_{k,1}^2 + c_{k,2}^2 \le -2c_{k,1}$ for $2 \le k \le n$. Hence it follows that, for $2 \le k \le n$,

$$\begin{aligned} |\lambda_{k,t}|^2 &= (1 + c_{k,1} a_t)^2 + c_{k,2}^2 a_t^2 \\ &\le 1 + 2c_{k,1} a_t - 2c_{k,1} a_t^2 \\ &= (1 + c_{k,1} a_t/2)^2 + c_{k,1} a_t - c_{k,1}^2 a_t^2/4 - 2c_{k,1} a_t^2. \end{aligned}$$

Since $-2 \le c_{k,1} < 0$, we have $c_{k,1} a_t - c_{k,1}^2 a_t^2/4 - 2c_{k,1} a_t^2 = |c_{k,1}| a_t (c_{k,1} a_t/4 + 2a_t - 1) \le 0$ for all $a_t \le 1/2$. Hence for all $t \ge T_2$ such that $a_t \le 1/2$, we have

$$|\lambda_{k,t}| = |\varphi(a_t, \omega_n^{k-1})| \le 1 + c_{k,1} a_t/2, \tag{30}$$

where $2 \le k \le n$. Denote $\delta^* = \inf_{2 \le k \le n} (1/2)|c_{k,1}| > 0$. Then it follows that

$$\Pi_{j=l}^t |\lambda_{k,j}| < \Pi_{j=l}^t (1 - \delta^* a_j) \tag{31}$$

for $2 \le k \le n$, where $t \ge l \ge T_2$. Hence we obtain

$$\begin{aligned} |M(a_t) \ldots M(a_k)\theta|^2 &= \theta^T F_n^* [\Pi_{j=k}^t \operatorname{diag}(0, \lambda_{2,j}, \ldots, \lambda_{n,j})]^* F_n F_n^* \\ &\quad \times [\Pi_{j=k}^t \operatorname{diag}(0, \lambda_{2,j}, \ldots, \lambda_{n,j})] F_n \theta \\ &\le \Pi_{j=k}^t (1 - \delta^* a_j)^2 |\theta|^2. \end{aligned}$$

This completes the proof. □

[2] When $a < 1$, the n diagonal entries of $M(a)$ are all positive, which ensures aperiodicity of $M(a)$.

Corollary 16. *Let θ, T_2 and δ^* be given as in Lemma* 15 *and denote $M(t,k) = M(a_t)\dots M(a_k)$ for $t > k \geq T_2$. Then $M^o(t,k) = F_n^*[\Pi_{j=k}^t \operatorname{diag}(0, \lambda_{2,j}, \dots, \lambda_{n,j})]F_n$ is a real matrix satisfying*

$$M(t,k)\theta = M^o(t,k)\theta. \tag{32}$$

Moreover, $|M^o(t,k)|_\infty \leq C\Pi_{j=k}^t(1-\delta^* a_j)$ for some $C > 0$ independent of t, k. The infinity norm $|\cdot|_\infty$ denotes the largest absolute value of the elements in the matrix.

Proof. Obviously $M^o(t,k)$ is a real matrix since $M^o(t,k) = M(a_t)\dots M(a_k) - (1/n)1_n 1_n^T$, and (32) follows from (28). The estimate for $|M^o(t,k)|_\infty$ follows from (31). □

Theorem 17. *Assume* (A1)–(A3). *Then algorithm* (2) *achieves* (i) *mean square consensus, and* (ii) *strong consensus for* (a) $\gamma \in (1/2, 1)$ *associated with any* $\alpha > 0$ *in* (A2), *and* (b) $\gamma = 1$ *provided that* $\alpha > 1/(2\delta^*)$.

Proof. The theorem is proved using the same procedure as in the two-agent case. For $\{\xi_t, t \geq 0\}$, we first write the recursion of ξ_t by (27) with the initial time $t = T_1 \vee T_2$ and show its mean square convergence by Lemma 13 and Lemma 15(ii). For proving almost sure convergence of ξ_t, we use Lemma 13, Corollary 16, and Lemma 9 to carry out the double array analysis, where we need to take $\alpha > 1/(2\delta^*)$ for the case $\gamma = 1$.

These combined with Lemma 12 lead to the mean square and almost sure convergence of the n sequences $\{x_t^i, t \geq 0\}$, $i \in \mathcal{N}$, to the same limit. □

For deterministic models, if the coefficient matrix in the consensus algorithm is doubly stochastic, the sum of the individual states remains a constant during the iterates. Moreover, if the algorithms achieve consensus, the state of each agent converges to the initial state average, giving the so-called average-consensus [34, 51]. In our model, due to the noise, the limit is a random variable differing from the initial state average although $M(a_t)$ is a doubly stochastic matrix. We have the following performance estimate which illustrates the effect of the noise.

Proposition 18. *Under* (A1)–(A3), *the state iterates in* (2) *satisfy*

$$E|\lim_{t\to\infty} x_t^i - \operatorname{ave}(x_0)| = \lim_{t\to\infty} E|x_t^i - \operatorname{ave}(x_0)|^2 = O(Q) \qquad \text{for all } i \in \mathcal{N}, \tag{33}$$

where $\operatorname{ave}(x_0) = (1/n)\sum_{k=1}^n x_0^k$ *is the initial state average and Q is the variance of the i.i.d. noises.*

Proof. This follows from the mean square consensus result in Theorem 17 and the relation (19). □

As the noise variance tends to zero, (33) indicates that the mean square error between $\lim_{t\to\infty} x_t^i$ and $\operatorname{ave}(x_0)$ converges to zero. This is consistent with the corresponding average-consensus results in deterministic models.

5. Consensus seeking on connected undirected graphs. In this section we consider more general network topologies but require that all links are bidirectional; i.e., we restrict our attention to undirected graphs.

Let the location of the n agents be associated with an undirected graph (to be simply called a graph) $G = (\mathcal{N}, \mathcal{E})$ consisting of a set of nodes $\mathcal{N} = \{1, 2, \dots, n\}$ and a set of edges $\mathcal{E} \subset \mathcal{N} \times \mathcal{N}$. We denote each edge as an unordered pair (i,j), where $i \neq j$. A path in G consists of a sequence of nodes $i_1, i_2, \dots, i_l$, $l \geq 2$, such that $(i_k, i_{k+1}) \in \mathcal{E}$ for all $1 \leq k \leq l-1$. The graph G is said to be connected if there exists a path between any two distinct nodes. The agent A_k (resp., node k) is a neighbor of A_i (resp., node i) if $(k,i) \in \mathcal{E}$, where $k \neq i$. Denote the neighbors of node i by $\mathcal{N}_i \subset \mathcal{N}$. We make the following assumption.

(A4) The undirected graph G is connected.

5.1. The measurement model and stochastic approximation. The formulation in section 2 is adapted to the undirected graph $G = (\mathcal{N}, \mathcal{E})$ as follows. For each $i \in \mathcal{N}$, we denote the measurement by agent A_i of agent A_k's state by

$$y_t^{ik} = x_t^k + w_t^{ik}, \qquad t \in \mathbb{Z}^+, \quad k \in \mathcal{N}_i, \tag{34}$$

where w_t^{ik} is the additive noise. Write the state vector $x_t = [x_t^1, \ldots, x_t^n]^T$. We introduce the following assumption which is slightly weaker for the noise condition than (A1).

(A1′) The noises $\{w_t^{ik}, t \in \mathbb{Z}^+, i \in \mathcal{N}, k \in \mathcal{N}_i\}$ are independent with respect to the indices i, k, t and also independent of x_0, and each w_t^{ik} has zero mean and variance $Q_t^{i,k}$. In addition, $E|x_0|^2 < \infty$ and $\sup_{t \ge 0, i \in \mathcal{N}} \sup_{k \in \mathcal{N}_i} Q_t^{ik} < \infty$.

We use the state updating rule

$$x_{t+1}^i = (1 - a_t)x_t^i + \frac{a_t}{|\mathcal{N}_i|} \sum_{k \in \mathcal{N}_i} y_t^{ik}, \qquad t \in \mathbb{Z}^+, \tag{35}$$

where $i \in \mathcal{N}$ and $a_t \in [0, 1]$, and we have the relation

$$x_{t+1}^i = x_t^i + a_t(m_t^i - x_t^i), \tag{36}$$

where $m_t^i = (1/|\mathcal{N}_i|) \sum_{k \in \mathcal{N}_i} y_t^{ik}$.

5.2. Stochastic Lyapunov functions. The specification of the stochastic Lyapunov function makes use of the relative positions of the agents. For agent A_i, we define its local potential as

$$P_i(t) = (1/2) \sum_{j \in \mathcal{N}_i} |x_t^i - x_t^j|^2, \qquad t \ge 0.$$

Accordingly, the total potential and total mean potential are given by

$$P_{\mathcal{N}}(t) = \sum_{i \in \mathcal{N}} P_i(t), \qquad V(t) = E \sum_{i \in \mathcal{N}} P_i(t), \quad t \ge 0.$$

It is easy to show that $m_t^i - x_t^i$ in (36) may be decomposed into the form

$$m_t^i - x_t^i = -\frac{1}{|\mathcal{N}_i|} \frac{\partial P_i(t)}{\partial x_t^i} + \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} w_t^{ij}. \tag{37}$$

This means the state of each agent is updated along the descent direction of the local potential subject to an additive noise, and justifies a stochastic approximation interpretation of algorithm (35). This interpretation is also applicable to digraphs.

Under (A4), it is easy to show that $P_{\mathcal{N}}(t) = 0$ if and only if $x_t^1 = \cdots = x_t^n$. For our convergence analysis, we will use $P_{\mathcal{N}}(t)$ as a stochastic Lyapunov function. We introduce the graph Laplacian for G as a symmetric matrix $L = (a_{ij})_{1 \le i,j \le n}$, where

$$a_{ij} = \begin{cases} d_i & \text{if } j = i, \\ -1 & \text{if } j \in \mathcal{N}_i, \\ 0 & \text{otherwise,} \end{cases} \tag{38}$$

and $d_i = |\mathcal{N}_i|$ is the degree (i.e., the number of neighbors) of node i. Denote $1_n = [1, 1, \ldots, 1]^T \in \mathbb{R}^n$. Since G is connected, $\text{rank}(L) = n - 1$ and the null space of L

is span$\{1_n\}$ [19, 35]. We have the following relation in terms of the graph Laplacian [19]:

$$P_{\mathcal{N}}(t) = (1/2)\sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}_i} |x_t^i - x_t^j|^2 = x_t^T L x_t, \qquad t \geq 0.$$

By (35), we have the state updating rule

$$x_{t+1}^i = (1-a_t)x_t^i + (a_t/|\mathcal{N}_i|)\sum_{j\in\mathcal{N}_i} x_t^j + (a_t/|\mathcal{N}_i|)\sum_{j\in\mathcal{N}_i} w_t^{ij}. \tag{39}$$

Denote

$$\tilde{w}_t^i = (1/|\mathcal{N}_i|)\sum_{j\in\mathcal{N}_i} w_t^{ij}, \qquad \tilde{w}_t = [\tilde{w}_t^1, \dots, \tilde{w}_t^n]^T. \tag{40}$$

With $d_i = |\mathcal{N}_i|$, we further introduce the matrix $\hat{L} = (\hat{a}_{ij})_{1\leq i,j\leq n}$, where

$$\hat{a}_{ij} = \begin{cases} 1 & \text{if } j = i, \\ -d_i^{-1} & \text{if } j \in \mathcal{N}_i, \\ 0 & \text{otherwise.} \end{cases} \tag{41}$$

Define the diagonal matrix $D_{\mathcal{N}} = \operatorname{diag}(d_1^{-1}, \dots, d_n^{-1})$. Note that $\hat{L} = D_{\mathcal{N}}L$.

LEMMA 19. *For $t \geq 0$ and $\{x_t, t \geq 0\}$ generated by* (34)–(35), *we have*

$$\begin{aligned} P_{\mathcal{N}}(t+1) = P_{\mathcal{N}}(t) &- 2a_t x_t^T L D_{\mathcal{N}} L x_t + a_t^2 x_t^T L D_{\mathcal{N}} L D_{\mathcal{N}} L x_t \\ &+ 2a_t x_t^T L\tilde{w}_t - 2a_t^2 x_t^T L D_{\mathcal{N}} L \tilde{w}_t + a_t^2 \tilde{w}_t^T L \tilde{w}_t. \end{aligned} \tag{42}$$

Proof. By (39), we get the vector equation

$$x_{t+1} = x_t - a_t\hat{L}x_t + a_t\tilde{w}_t, \qquad t \geq 0, \tag{43}$$

which leads to the recursion of the total potential as follows:

$$\begin{aligned} P_{\mathcal{N}}(t+1) &= x_{t+1}^T L x_{t+1} \\ &= x_t^T L x_t - 2a_t x_t^T L D_{\mathcal{N}} L x_t + a_t^2 x_t^T L D_{\mathcal{N}} L D_{\mathcal{N}} L x_t \\ &\quad + 2a_t x_t^T L \tilde{w}_t - 2a_t^2 x_t^T L D_{\mathcal{N}} L \tilde{w}_t + a_t^2 \tilde{w}_t^T L \tilde{w}_t, \end{aligned}$$

and the lemma follows. □

In the subsequent proofs, we use $A \Rightarrow B$ as the abbreviation for "A implies B," and $A \Leftrightarrow B$ for "A is equivalent to B."

LEMMA 20. *Under* (A4), *we have the following assertions:*

(i) *The null spaces of L, $LD_{\mathcal{N}}L$, and $LD_{\mathcal{N}}LD_{\mathcal{N}}L$ are each given by* span$\{1_n\}$.

(ii) *There exist $c_1 > 0$ and $c_2 > 0$ such that $LD_{\mathcal{N}}L \geq c_1 L$ and $LD_{\mathcal{N}}LD_{\mathcal{N}}L \leq c_2 L$.*

(iii) *In addition, we assume* (A1′)–(A2′) *and let T_c be such that $1-2a_tc_1+a_t^2c_2 \geq 0$ for all $t \geq T_c$. Then for all $t \geq T_c$, we have*

$$V(t+1) \leq (1 - 2a_tc_1 + a_t^2c_2)V(t) + O(a_t^2). \tag{44}$$

Proof. See the appendix. □

THEOREM 21. *Under* (A1′)–(A2′) *and* (A4), *algorithm* (35) *achieves weak consensus.*

Proof. For T_c given in Lemma 20(iii), we select $\hat{T}_c \ge T_c$ to ensure $a_t \le c_1 c_2^{-1}$. Hence $1 - c_1 a_t \ge 1 - 2c_1 a_t + c_2 a_t^2 \ge 0$ for all $t \ge \hat{T}_c$, and we find a fixed constant $C > 0$ such that

$$V(t+1) \le (1 - c_1 a_t)V(t) + C a_t^2$$

for all $t \ge \hat{T}_c$; this leads to

$$V(t+1) \le V(\hat{T}_c) \prod_{j=\hat{T}_c}^{t} (1 - c_1 a_j) + C \sum_{k=\hat{T}_c}^{t} \prod_{j=k+1}^{t} (1 - c_1 a_j) a_k^2, \tag{45}$$

where $\prod_{j=t+1}^{t}(1 - c_1 a_j) \triangleq 1$. Under (A2′), elementary estimates using (45) yield

$$\lim_{t\to\infty} V(t) = 0. \tag{46}$$

It immediately follows that

$$\lim_{t\to\infty} E|x_t^i - x_t^k|^2 = 0, \qquad i \in \mathcal{N},\ k \in \mathcal{N}_i. \tag{47}$$

Since G is connected, there exists a path between any pair of distinct nodes i and k. By repeatedly applying (47) to all pairs of neighboring nodes along that path, we can show that $\lim_{t\to\infty} E|x_t^i - x_t^k|^2 = 0$ for any $i, k \in \mathcal{N}$. □

COROLLARY 22. *In Theorem* 21*, we assume all other assumptions but replace* (A2′)(ii) *by the condition* (H)*: There exists* $T_0 > 0$ *such that for* $t \ge T_0$, $\alpha_0 t^{-\gamma_0} \le a_t \le \beta_0 t^{-\gamma_0}$ *holds for some* $0 < \alpha_0 < \beta_0 < \infty$ *and* $\gamma_0 \in (0, 1/2]$*. Then algorithm* (35) *still achieves weak consensus.*

Proof. For (45), we have $\prod_{j=k+1}^{t}(1 - c_1 a_j)a_k^2 \le \prod_{j=k+1}^{t}(1 - c_1 a_j/2)^2 a_k^2$. We apply Corollary 7 to show that (46) still holds. This completes the proof. □

Remark. Notice that under (H), $\sum_{t=0}^{\infty} a_t^2 = \infty$. The conditions of Corollary 22 in general do not ensure mean square consensus.

5.3. The direction of invariance.

Theorem 21 shows the difference between the states of any two agents converges to zero in mean square. However, this alone does not mean that they will converge to a common limit. The asymptotic vanishing of the stochastic Lyapunov function indicates only that the state vector x_t will approach the subspace $\mathrm{span}\{1_n\}$. To obtain mean square consensus results, we need some additional estimation. The strategy is to show that the oscillation of the sequence $\{x_t, t \ge 0\}$ along the direction 1_n will gradually die off. This is achieved by proving the existence of a vector η which is not orthogonal to 1_n and such that the linear combination $\eta^T x_t$ of the components in x_t converges. For convenience, η will be chosen to satisfy the additional requirement that $\eta^T x_{t+1}$ depends not on the whole of x_t but only on $\eta^T x_t$; this will greatly facilitate the associated calculation.

DEFINITION 23. *Let* $x_t = [x_t^1, \ldots, x_t^n]^T$ *be generated by* (35)*. If* $\eta = [\eta_1, \ldots, \eta_n]^T$ *is a real-valued vector of unit length, i.e.,* $|\eta|^2 = \sum_{i=1}^{n} \eta_i^2 = 1$*, and satisfies*

$$\eta^T x_{t+1} = \eta^T x_t + a_t \eta^T \tilde{w}_t, \qquad t \ge 0, \tag{48}$$

for any initial condition x_0 *and any step size sequence* $a_t \in [0, 1]$*, where* $\tilde{w}_t$ *is given in* (40)*, then* η *is called a direction of invariance associated with* (35).

The directions of invariance associated with the consensus algorithm (35) are easily characterized in terms of the degrees of the nodes of the underlying graph.

Theorem 24. *We have the following assertions:*

(i) *There exists a real-valued vector $\eta = (\eta_1, \dots, \eta_n)^T$ of unit length satisfying $\eta^T \hat{L} = 0$, where $\hat{L}$ is defined by* (41).

(ii) *If $|\eta| = 1$, then η is a direction of invariance for* (35) *if and only if $\eta^T \hat{L} = 0$.*

(iii) *Under* (A4)*, the direction of invariance for* (35) *has the representation $\eta = c[d_1, \dots, d_n]^T$, where $c = \pm(\sum_{i=1}^n d_i^2)^{1/2}$ and $d_i = |\mathcal{N}_i|$ is the degree of node i.*

Proof. It is easy to prove (i) since $\hat{L}$ does not have full rank, and η is in fact the left eigenvector of $\hat{L}$ associated with the eigenvalue 0.

We now show (ii). The condition $\eta^T \hat{L} = 0$ combined with (43) implies

$$\eta^T x_{t+1} = \eta^T x_t + a_t \eta^T \tilde{w}_t.$$

The sufficiency part of (ii) follows easily. Conversely, if the unit length vector η satisfies (48) for all initial states x_0^i and the step size a_t as specified in Definition 23, then we necessarily have $\eta^T \hat{L} = 0$. So the necessity part of (ii) holds.

We continue to prove (iii) under (A4). By (ii) and the definition of $\hat{L}$, η with $|\eta| = 1$ is a direction of invariance if and only if $\eta^T D_{\mathcal{N}} L = 0$, which in turn is equivalent to $L D_{\mathcal{N}} \eta = 0$. By (A4) and Lemma 20, we have $D_{\mathcal{N}} \eta = c 1_n$, where $c \neq 0$ is a constant to be determined. This gives $\eta = c[d_1, \dots, d_n]^T$, where c is determined by the condition $|\eta| = 1$. The direction of invariance is unique up to sign. □

If η is a direction of invariance, then Theorem 24 shows under (A4) that all elements of η have the same sign. Therefore, η is not orthogonal to 1_n, and the requirement stated at the beginning of this section is met. Geometrically, the notion of the direction of invariance means under (35) and zero noise conditions, the projection (i.e., $(\eta^T x_t)\eta$) of x_t in $\mathbb{R}^n$ along the direction η would remain a constant vector regardless of the value of $a_t \in [0, 1]$ used in the iterates.

5.4. Mean square consensus. Now we are in a position to establish mean square consensus.

Lemma 25. *Assume* (A1′)*–*(A2′) *and* (A4)*, and let $\{x_t, t \geq 0\}$ be given by* (35)*, $\eta_0 = [d_1, \dots, d_n]^T$, where $d_i = |\mathcal{N}_i|$. Then there exists a random variable y^* such that $\lim_{t\to\infty} E|\eta_0 x_t - y^*|^2 = 0$.*

Proof. By Theorem 24, $\eta_0/|\eta_0|$ is a direction of invariance. Hence, we have

$$\eta_0^T x_{t+1} = \eta_0^T x_0 + a_0 \eta_0^T \tilde{w}_0 + \dots + a_t \eta_0^T \tilde{w}_t.$$

By (A1′) and (A2′), it follows that $\eta_0^T x_t$ converges in mean square. □

The weak consensus result combined with the convergence of $\eta_0^T x_t$ ensures that x_t itself converges.

Theorem 26. *Under* (A1′)–(A2′) *and* (A4)*, algorithm* (35) *achieves mean square consensus.*

Proof. By Theorem 21, we have weak consensus, i.e.,

$$\lim_{t\to\infty} E|x_t^i - x_t^k|^2 = 0 \qquad \text{for all } i, k \in \mathcal{N}. \tag{49}$$

On the other hand, by Lemma 25, as $t \to \infty$,

$$\eta_0^T x_t = \eta_0^T [x_t^1 - x_t^1, \dots, x_t^n - x_t^1]^T + \eta_0^T [x_t^1, \dots, x_t^1]^T$$

converges in mean square, which combined with (49) implies x_t^1 converges in mean square. By (49) again, the mean square consensus result follows. □

6. Leader following and convergence. Now we apply the stochastic Lyapunov function approach to the scenario of leader following [23, 44]. Suppose there are n agents located in the digraph $G_d = (\mathcal{N}, \mathcal{E})$, and without loss of generality, denote the leader by agent A_1. We denote by $\mathcal{N}_F = \mathcal{N}\backslash\{1\}$ the set of follower agents. For $i \in \mathcal{N}$, denote the individual states by x_t^i, $t \in \mathbb{Z}^+$. The leader A_1 does not receive measurements from other agents; to capture this feature in G_d, there is no edge reaching A_1 from other agents. The initial state of A_1 is chosen randomly, after which the state remains constant. That is, $x_t^1 \equiv \vartheta$, where ϑ is a random variable, which is unknown to any other agent A_i, $i \in \mathcal{N}_F$.

For node $i \in \mathcal{N}_F$, its measurement is given as

$$y_t^{i,k} = x_t^k + w_t^{i,k}, \qquad t \in \mathbb{Z}^+, \quad k \in \mathcal{N}_i,$$

where $w_t^{i,k}$ is the additive noise. For $i \in \mathcal{N}_F$, the state is updated by

$$x_{t+1}^i = (1 - a_t)x_t^i + \frac{a_t}{|\mathcal{N}_i|} \sum_{j\in\mathcal{N}_i} y_t^{ij}. \tag{50}$$

We adapt (A1′) to the graph $G_d = (\mathcal{N}, \mathcal{E})$ in an obvious manner. But it should be kept in mind that in this leader following model the noise term w_t^{ik} is defined only for $i \in \mathcal{N}_F$ since the leader has no neighbor. Also, $x_0^1 \equiv \vartheta$ since A_1 is the leader, and under (A1′), we have $E|\vartheta|^2 < \infty$.

To make the problem nontrivial, we use the following underlying assumption.

(A5) In $G_d = (\mathcal{N}, \mathcal{E})$, node 1 is the neighbor of at least one node in $\mathcal{N}_F$.

Now, based on the digraph $G_d = (\mathcal{N}, \mathcal{E})$, we set each $(i, j) \in \mathcal{E}$ as an unordered pair and this procedure induces an undirected graph $G_u = (\mathcal{N}, \mathcal{E}^u)$ with its associated graph Laplacian L^u. We decompose L^u into the form

$$L^u = \begin{bmatrix} L_1^u \\ L_{n-1}^u \end{bmatrix},$$

where L_1^u is the first row in L^u.

In order to develop the stochastic Lyapunov analysis, we need some restrictions on the set of nodes $\mathcal{N}_F$ and the associated edges. Let $(\mathcal{N}_F, \mathcal{E}_F)$ denote the directed subgraph of $(\mathcal{N}, \mathcal{E})$ obtained by removing node 1 and all edges containing 1 as the initial node. We introduce the following assumption.

(A6) An ordered pair $(i, j) \in \mathcal{E}_F$ implies the ordered pair (j, i) is also in $\mathcal{E}_F$.

Remark. (A5)–(A6) imply that at least one follower can receive information from the leader while the information exchange among the followers is bidirectional.

In analogy to the construction of G_u, we induce from the digraph $(\mathcal{N}_F, \mathcal{E}_F)$ an undirected graph, denoted by $G_{Fu} = (\mathcal{N}_F, \mathcal{E}_F^u)$. We introduce the following assumption.

(A7) The undirected graph $G_{Fu} = (\mathcal{N}_F, \mathcal{E}_F^u)$ is connected.

PROPOSITION 27. *Under* (A5)–(A7), *the undirected graph* $G_u = (\mathcal{N}, \mathcal{E}^u)$ *is connected and* $\mathrm{rank}(L^u) = \mathrm{rank}(L_{n-1}^u) = n - 1$.

Proof. It is obvious that G_u is connected. Hence $\mathrm{rank}(L^u) = n - 1$. Since $1_n^T L^u = 0$, it follows that L_1^u is a linear combination of the rows in L_{n-1}^u, which implies $\mathrm{rank}(L_{n-1}^u) = n - 1$. $\square$

Denote $x_{\vartheta,t} = [\vartheta, x_t^2, \dots, x_t^n]^T$, $\tilde{w}_i = (1/|N_i|) \sum_{k\in\mathcal{N}_i} w_t^{i,k}$ for $i \geq 2$, and $\tilde{w}_t = [0, \tilde{w}_t^2, \dots, \tilde{w}_t^n]^T$. Write $D_0 = \mathrm{diag}(0, d_2^{-1}, d_n^{-1})$. By writing (50) in the vector form, we get the following lemma.

LEMMA 28. *We have the recursion for the state vector*

$$x_{\vartheta,t+1} = x_{\vartheta,t} - a_t D_0 L^u x_{\vartheta,t} + a_t \tilde{w}_t, \qquad t \geq 0,$$

where $x_{\vartheta,t}$ is generated by algorithm (50).

THEOREM 29. *Under* (A1′)–(A2′), (A5)–(A7), *and algorithm* (50), *we have*

$$\lim_{t\to\infty} E|x_t^i - \vartheta|^2 = 0 \tag{51}$$

for all $i \in \mathcal{N}_F$, where ϑ is the fixed random variable as the state for the leader.

Proof. Step 1. Define the stochastic Lyapunov function $P_{\vartheta,\mathcal{N}}(t) = x_{\vartheta,t}^T L^u x_{\vartheta,t}$, where $L^u \geq 0$, and denote $V(t) = EP_{\vartheta,\mathcal{N}}(t)$, $t \geq 0$. By Lemma 28, it is easy to show

$$\begin{aligned} V(t+1) = V(t) &- 2a_t E[x_{\vartheta,t}^T L^u D_0 L^u x_{\vartheta,t}] \\ &+ a_t^2 E[x_{\vartheta,t}^T L^u D_0 L^u D_0 L^u x_{\vartheta,t}] + O(a_t^2). \end{aligned} \tag{52}$$

Let $y_\theta = [\theta, y_2, \ldots, y_n]^T$, where θ denotes a fixed real number. First, by $\mathrm{rank}(L^u) = n-1$, we can show that $L^u y_\theta = 0 \Leftrightarrow y_\theta = \theta 1_n^T$. Obviously $L^u y_\theta = 0 \Rightarrow L^u D_0 L^u y_\theta = 0$ $\Rightarrow L^u D_0 L^u D_0 L^u y_\theta = 0$. On the other hand, letting $L^u = [(L^u)^{1/2}]^2$, where $(L^u)^{1/2} \geq 0$, we have $L^u D_0 L^u D_0 L^u y_\theta = 0 \Rightarrow (L^u)^{1/2} D_0 L^u y_\theta = 0 \Rightarrow L^u D_0 L^u y_\theta = 0 \Rightarrow$ $\mathrm{diag}(0, d_2^{-1/2}, \ldots, d_n^{-1/2}) L^u y_\theta = 0 \Leftrightarrow \mathrm{diag}(d_2^{-1/2}, \ldots, d_n^{-1/2}) L_{n-1}^u y_\theta = 0 \Leftrightarrow L_{n-1}^u y_\theta = 0 \Leftrightarrow L^u y_\theta = 0$ since $\mathrm{rank}(L_{n-1}^u) = \mathrm{rank}(L^u) = n-1$ by Proposition 27. Now we conclude that $\theta 1_n$ is the unique point where each of $y_\theta^T L^u y_\theta$, $y_\theta^T L^u D_0 L^u y_\theta$, and $y_\theta^T L^u D_0 L^u D_0 L^u y_\theta$ attains its minimum 0.

Step 2. Letting $y^{(n-1)} = [y_2, \ldots, y_n]^T$, we introduce three positive semidefinite quadratic forms in terms of $y^{(n-1)}$: $Q_1(y^{(n-1)}) = y_\theta^T L^u y_\theta$, $Q_2(y^{(n-1)}) = y_\theta^T L^u D_0 L^u y_\theta$, and $Q_3(y^{(n-1)}) = y_\theta^T L^u D_0 L^u D_0 L^u y_\theta$. Let $z = y^{(n-1)} - \theta 1_{n-1}$, and we may write

$$0 \leq Q_1(y^{(n-1)}) = z^T M_1 z + v^T z + c,$$

where M_1 is an $(n-1) \times (n-1)$ symmetric matrix, $v \in \mathbb{R}^{n-1}$, and $c \in \mathbb{R}$. Clearly $z^T M_1 z + v^T z + c = 0 \Leftrightarrow z = 0$ since $Q_1(y^{(n-1)}) = 0 \Leftrightarrow y_\theta = \theta 1_n$ by Step 1; by elementary linear algebra and a contradictory argument we can show $c = 0$, $v^T = 0$, and $M_1 > 0$. Hence, $Q_1(y^{(n-1)}) = z^T M_1 z$. Since M_1 is constructed based on the second order coefficient of $y^{(n-1)}$ in $y_\theta^T L^u y_\theta$, we see that M_1 is independent of θ. Similarly, we can find matrices $M_2 > 0$ and $M_3 > 0$, both independent of θ, such that

$$Q_2(y^{(n-1)}) = z^T M_2 z, \qquad Q_3(y^{(n-1)}) = z^T M_3 z,$$

where $z = y^{(n-1)} - \theta 1_{n-1}$. We denote the smallest and largest eigenvalue of M_i, respectively, by $\lambda_{i,\min} > 0$ and $\lambda_{i,\max} > 0$ for $i = 1, 2, 3$. Now we have

$$Q_2(y^{(n-1)}) = z^T M_2 z \geq \lambda_{2,\min} \lambda_{1,\max}^{-1} z^T M_1 z = \lambda_{2,\min} \lambda_{1,\max}^{-1} Q_1(y^{(n-1)}), \tag{53}$$

$$Q_3(y^{(n-1)}) = z^T M_3 z \leq \lambda_{3,\max} \lambda_{1,\min}^{-1} z^T M_1 z = \lambda_{3,\max} \lambda_{1,\min}^{-1} Q_1(y^{(n-1)}). \tag{54}$$

Step 3. Now it follows from (52) and (53)–(54) that

$$V(t+1) \leq (1 - 2\tau_1 a_t + \tau_2 a_t^2) V(t) + O(a_t^2), \tag{55}$$

where $\tau_1 = \lambda_{2,\min} \lambda_{1,\max}^{-1}$ and $\tau_2 = \lambda_{3,\max} \lambda_{1,\min}^{-1}$. Consequently, by use of product estimates as in (45), we can show $\lim_{t\to\infty} V(t) = 0$. Since the first entry in $x_{\vartheta,t}$

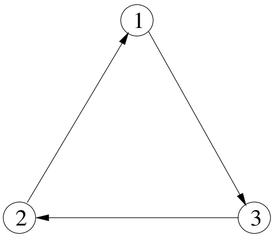


FIG. 4. *A digraph with* 3 *nodes.*

is ϑ and the associated undirected graph $G_u = (\mathcal{N}, \mathcal{E}^u)$ is connected, by the same argument as in proving weak consensus in Theorem 21, we can obtain (51). $\square$

Remark. In Theorem 29, if (A2′)(ii) is replaced by the condition (H): $\alpha_0 t^{-\gamma_0} \le a_t \le \beta_0 t^{-\gamma_0}$ for $t \ge T_0$, where $\alpha_0 > 0$ and $\gamma_0 \in (0, 1/2]$ (see Corollary 22), then Theorem 29 still holds. This may be proved by combining the proving argument for Corollary 22 with (55) to get $\lim_{t\to\infty} V(t) = 0$.

7. Numerical studies.

7.1. Simulations with a symmetric digraph. The digraph is shown in Figure 4, where $\mathcal{N}_1 = \{2\}, \mathcal{N}_2 = \{3\}$, and $\mathcal{N}_3 = \{1\}$. The initial condition for $x_t = [x_t^1, x_t^2, x_t^3]$ is $[4, 3, 1]$ at $t = 0$, and the i.i.d. Gaussian measurement noises have variance $\sigma^2 = 0.01$. Figure 5 shows the simulation with equal weights to an agent's neighbors and itself (as in Example 1) in the averaging rule ($x_{t+1}^1 = (x_t^1 + y_t^{12})/2$, etc.), without obtaining consensus. Figure 6 shows the convergence of algorithm (2) with the step size sequence $\{a_t = (t+5)^{-0.85}, t \ge 0\}$.

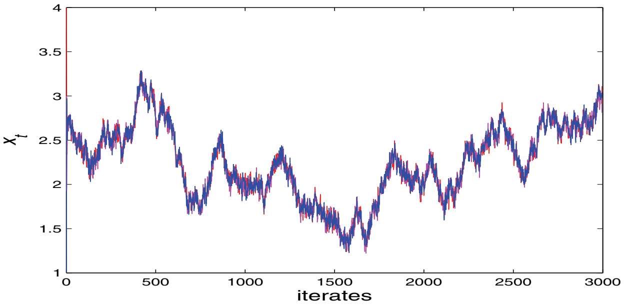


FIG. 5. *Equal weights are used for each agent's state and observation.*

7.2. Simulations with an undirected graph. The undirected graph is shown in Figure 7 with $\mathcal{N} = \{1, 2, 3, 4\}$ and $\mathcal{E} = \{(1, 2), (2, 3), (2, 4)\}$. The initial condition is $x_t|_{t=0} = [5, 1, 3, 2]^T$, and the i.i.d. Gaussian noises have variance $\sigma^2 = 0.01$. The simulation of the averaging rule with equal weights is given in Figure 8; hence we have $x_{t+1}^1 = (x_t^1 + y_t^{12})/2$ and $x_{t+1}^2 = x_t^2/4 + (y_t^{21} + y_t^{23} + y_t^{24})/4$, etc., where $t \ge 0$. It is seen that the 4 state trajectories in Figure 8 move towards each other rather quickly at the beginning, but they maintain long term fluctuations as the state iteration continues. The stochastic algorithm (35) is used in Figure 9, where $a_t = (t+5)^{-0.85}$, $t \ge 0$. Figure 9 shows the 4 trajectories all converge to the same constant level.

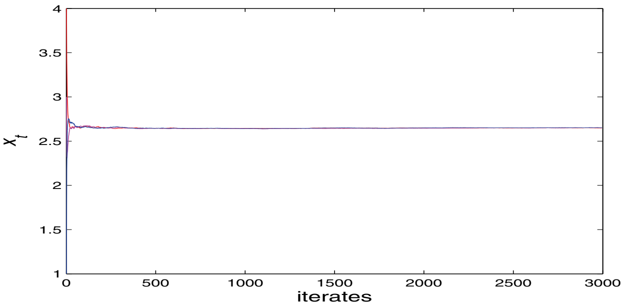


FIG. 6. *The* 3*-agent example using decreasing step size* $a_t = (t+5)^{-0.85}$*.*

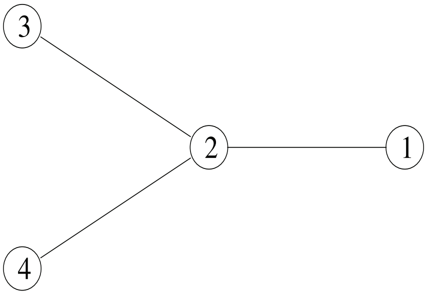


FIG. 7. *The undirected graph with* 4 *nodes.*

7.3. The leader following model. We adapt the undirected graph in Figure 7 to the leader following situation as follows. We set node 1 as the leader (without a neighbor) and $\mathcal{N}_2 = \{1, 3, 4\}$, $\mathcal{N}_3 = \{2\}$, $\mathcal{N}_4 = \{2\}$. We take $x_0^1 \equiv 4$ and the initial condition is given as $x_t|_{t=0} = [4, 2, 1, 3]^T$. Figure 10 shows the simulation with equal weights for each follower agent and its neighbors. We see that all three states of the follower agents move into a neighborhood of the constant level 4 and oscillate around that value. Compared with Figure 8, the trajectories of the followers in Figure 10 have a far smaller fluctuation. The reason is that in the leader following case, the total potential attains its minimum only at the leader's state rather than at all points in span$\{1_n\}$, which results in more regular behavior for the agents. In Figure 11 we show the simulation of algorithm (50) with $a_t = (t+5)^{-0.65}$, $t \geq 0$, which exhibits a satisfactory convergent behavior.

8. Concluding remarks. We consider consensus problems for networked agents with noisy measurements. First, the double array analysis is developed to analyze mean square and almost sure convergence. Next, stochastic Lyapunov functions are introduced to prove mean square consensus with the aid of the so-called direction of invariance, and this approach is further applied to leader following. We note that the methods developed in this paper may be extended to deal with general digraphs, and the second order moment condition for the noise may be relaxed when applying the stochastic double array analysis; see the recent work [22] for details. For future work, it is of interest to develop stochastic algorithms in models with dynamic topologies and asynchronous state updates, and in particular, extend the double array analysis to networks with switching topologies.

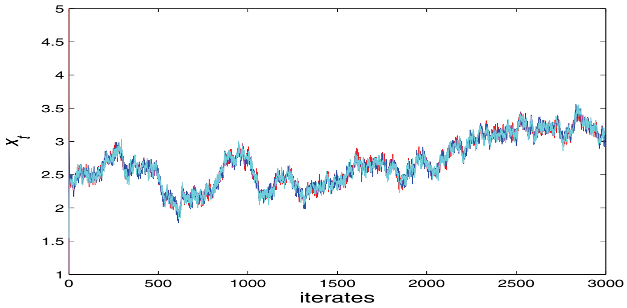


Fig. 8. *The 4-agent example using equal weights for each agent's state and observations.*

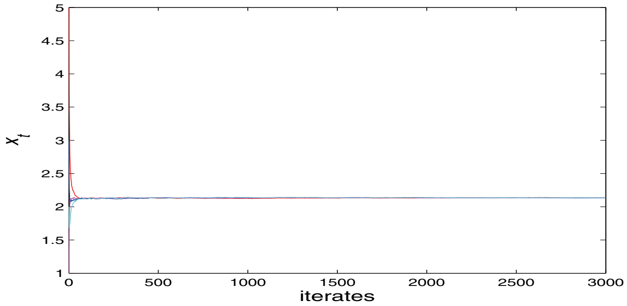


Fig. 9. *The 4-agent example using a decreasing step size* $a_t = (t+5)^{-0.85}$.

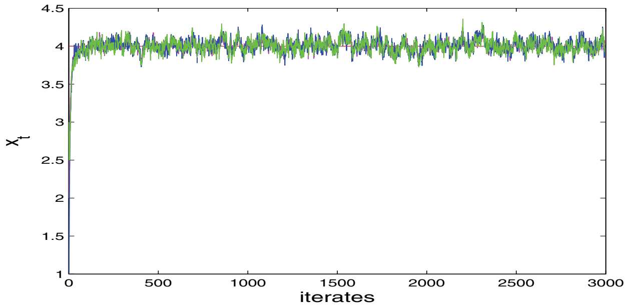


Fig. 10. *Leader following using equal weights for each follower agent's state and observations.*

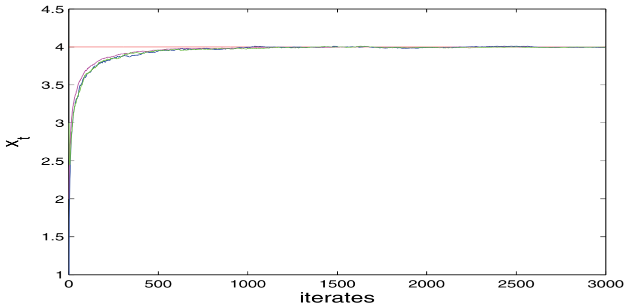


FIG. 11. *Leader following using a decreasing step size* $a_t = (t+5)^{-0.65}$.

Appendix.

Proof of Lemma 6. For case (i), by (10) we have

$$\sum_{k=T_1}^{t} \Pi_{t,k}^2 \le \sum_{k=T_1}^{t} \frac{\beta^2 (k+1)^{4\alpha}}{k^2 (t+1)^{4\alpha}}.$$

The desired upper bound is obtained from elementary estimates by considering three scenarios for α as in (14).

We continue with the estimate for case (ii). Let $\delta = 2\alpha/(1-\gamma) > 0$ and define

$$S_t = \sum_{k=1}^{t} k^{-2\gamma} e^{2\delta(k+1)^{1-\gamma}}, \qquad H_t = t^{-\gamma} e^{2\delta(t+1)^{1-\gamma}}, \quad t \ge 1.$$

Clearly there exists a sufficiently large $t_0 > 0$ such that H_t is strictly increasing for $t \ge t_0$. In addition, both S_t and H_t diverge to infinity. If we can show that for $t > t_0$,

$$0 < R_t = \frac{S_t - S_{t-1}}{H_t - H_{t-1}} \to R^*, \qquad \text{as } t \to \infty, \tag{A.1}$$

for some $R^* > 0$, then it is straightforward to show that $S_t = O(H_t)$. To show the existence of a limit in (A.1), we write

$$R_t = \frac{t^{-2\gamma} e^{2\delta(t+1)^{1-\gamma}}}{t^{-\gamma} e^{2\delta(t+1)^{1-\gamma}} - (t-1)^{-\gamma} e^{2\delta t^{1-\gamma}}}. \tag{A.2}$$

We have

$$\begin{aligned}
& t^{-\gamma} e^{2\delta(t+1)^{1-\gamma}} - (t-1)^{-\gamma} e^{2\delta t^{1-\gamma}} \\
&= t^{-\gamma} e^{2\delta(t+1)^{1-\gamma}} - t^{-\gamma} e^{2\delta[(t+1)^{1-\gamma} + t^{1-\gamma} - (t+1)^{1-\gamma}]} (1 - t^{-1})^{-\gamma} \\
&= t^{-\gamma} e^{2\delta(t+1)^{1-\gamma}} [1 - e^{2\delta[t^{1-\gamma} - (t+1)^{1-\gamma}]} (1 - t^{-1})^{-\gamma}] \\
&= t^{-\gamma} e^{2\delta(t+1)^{1-\gamma}} \left[1 - e^{-2\delta[(1-\gamma)t^{-\gamma} + o(t^{-\gamma})]}\right] [1 + \gamma t^{-1} + o(t^{-1})] \\
&= t^{-\gamma} e^{2\delta(t+1)^{1-\gamma}} [2\delta(1-\gamma) t^{-\gamma} + o(t^{-\gamma})][1 + \gamma t^{-1} + o(t^{-1})] \\
&= 2\delta(1-\gamma) t^{-2\gamma} e^{2\delta(t+1)^{1-\gamma}} [1 + o(1)] \\
&= 4\alpha t^{-2\gamma} e^{2\delta(t+1)^{1-\gamma}} [1 + o(1)]. \qquad \text{(A.3)}
\end{aligned}$$

By combining (A.2) and (A.3), it follows that $\lim_{t\to\infty} R_t = 4\alpha > 0$, and hence $S_t = O(H_t)$. Subsequently, we have

$$\begin{aligned}\sum_{k=T_1}^{t} \Pi_{t,k}^2 &= O\left(e^{-2\delta(t+1)^{1-\gamma}} \sum_{k=1}^{t} k^{-2\gamma} e^{2\delta(k+1)^{1-\gamma}}\right)\\ &= O(e^{-2\delta(t+1)^{1-\gamma}} H_t)\\ &= O(t^{-\gamma}),\end{aligned}$$

which completes the proof for case (ii), and the lemma follows. □

Proof of Lemma 20. (i) First, it is a well-known fact [19, 35] that when G is connected, the null space of L is $\operatorname{span}\{1_n\}$. Since $L \geq 0$, there exists a positive semidefinite matrix, denoted as $L^{1/2}$ such that $L = (L^{1/2})^2$. We also write $D_{\mathcal{N}}^{1/2} = \operatorname{diag}(d_1^{-1/2}, \dots, d_n^{-1/2})$ which gives $D_{\mathcal{N}} = (D_{\mathcal{N}}^{1/2})^2$. For $x \in \mathbb{R}^n$, we have $Lx = 0 \Rightarrow LD_{\mathcal{N}}Lx = 0 \Rightarrow LD_{\mathcal{N}}LD_{\mathcal{N}}Lx = 0$.

On the other hand, we have

$$\begin{aligned}LD_{\mathcal{N}}LD_{\mathcal{N}}Lx = 0 &\Rightarrow x^T LD_{\mathcal{N}}LD_{\mathcal{N}}Lx = 0\\ &\Leftrightarrow |L^{1/2}D_{\mathcal{N}}Lx|^2 = 0 \Leftrightarrow L^{1/2}D_{\mathcal{N}}Lx = 0\\ &\Rightarrow LD_{\mathcal{N}}Lx = 0 \Rightarrow x^T LD_{\mathcal{N}}Lx = 0\\ &\Leftrightarrow D_{\mathcal{N}}^{1/2}Lx = 0 \Leftrightarrow Lx = 0.\end{aligned}$$

Hence, it follows that $Lx = 0 \Leftrightarrow LD_{\mathcal{N}}Lx = 0 \Leftrightarrow LD_{\mathcal{N}}LD_{\mathcal{N}}Lx = 0$, and assertion (i) follows. Hence the matrices L, $LD_{\mathcal{N}}L$, and $LD_{\mathcal{N}}LD_{\mathcal{N}}L$ each have a rank of $n-1$.

(ii) We begin by proving the first part. Let $0 = \lambda_1,\ 0 < \lambda_2 \leq \lambda_3 \leq \cdots \leq \lambda_n$ and $0 = \hat\lambda_1,\ 0 < \hat\lambda_2 \leq \hat\lambda_3 \leq \cdots \leq \hat\lambda_n$, respectively, denote the eigenvalues of L and $LD_{\mathcal{N}}L$. Let $\Phi = (\alpha_1, \dots, \alpha_n)$ and $\hat\Phi = (\hat\alpha_1, \dots, \hat\alpha_n)$ be two orthogonal matrices (i.e., $\Phi^T\Phi = I$, and $\hat\Phi^T\hat\Phi = I$) such that

$$L\Phi = \Phi \operatorname{diag}(\lambda_1, \dots, \lambda_n), \qquad LD_{\mathcal{N}}L\hat\Phi = \hat\Phi \operatorname{diag}(\hat\lambda_1, \dots, \hat\lambda_n).$$

In view of $\lambda_1 = \hat\lambda_1 = 0$, we get $L\alpha_1 = LD_{\mathcal{N}}L\hat\alpha_1 = 0$. By (i), we necessarily have either $\alpha_1 = \hat\alpha_1$ or $\alpha_1 = -\hat\alpha_1$. In fact, we may take $\alpha_1 = \hat\alpha_1 = \pm(1/\sqrt{n}) \cdot 1_n$. Consequently, it is easy to show that $\operatorname{span}\{\alpha_2, \dots, \alpha_n\} = \operatorname{span}\{\hat\alpha_2, \dots, \hat\alpha_n\}$, which is the orthogonal complement of $\operatorname{span}\{1_n\}$ in $\mathbb{R}^n$.

Take any $x \in \mathbb{R}^n$. We may write $x = \sum_{i=1}^n y_i\alpha_i$, $x = \sum_{i=1}^n \hat y_i\hat\alpha_i$, where $y = (y_1, \dots, y_n)$, $\hat y = (\hat y_1, \dots, \hat y_n)$ are uniquely determined and satisfy $\sum_{i=1}^n y_i^2 = \sum_{i=1}^n \hat y_i^2 = |x|^2$. Recalling that we have taken $\alpha_1 = \hat\alpha_1 \neq 0$, it necessarily follows that $y_1 = \hat y_1$ since, otherwise, $(y_1 - \hat y_1)\alpha_1 \in \operatorname{span}\{\alpha_2, \dots, \alpha_n\}$ with $y_1 - \hat y_1 \neq 0$, which is impossible. Hence we get

$$\sum_{i=2}^n y_i^2 = \sum_{i=2}^n \hat y_i^2. \tag{A.4}$$

For $x \in \mathbb{R}^n$, since $\lambda_1 = \hat\lambda_1 = 0$ we have the estimate

$$x^T LD_{\mathcal{N}}Lx = \hat y^T \hat\Phi^T LD_{\mathcal{N}}L\hat\Phi\hat y = \sum_{i=2}^n \hat\lambda_i \hat y_i^2 \geq \hat\lambda_2 \sum_{i=2}^n \hat y_i^2.$$

On the other hand, we have $x^T Lx \leq \lambda_n \sum_{i=2}^n y_i^2 = \lambda_n \sum_{i=2}^n \hat{y}_i^2$, where the equality follows from (A.4). Hence it follows that $x^T LD_{\mathcal{N}} Lx \geq \hat{\lambda}_2 \lambda_n^{-1} x^T Lx$, and therefore, the first part of (ii) is proved by taking $c_1 = \hat{\lambda}_2 \lambda_n^{-1} > 0$.

We denote the eigenvalues of $LD_{\mathcal{N}} LD_{\mathcal{N}} L$ by $0 = \tilde{\lambda}_1,\ 0 < \tilde{\lambda}_2 \leq \tilde{\lambda}_3 \leq \cdots \leq \tilde{\lambda}_n$. Following a very similar argument, we can show that for any $x \in \mathbb{R}^n$,

$$x^T LD_{\mathcal{N}} LD_{\mathcal{N}} Lx \leq \tilde{\lambda}_n \lambda_2^{-1} x^T Lx,$$

which implies the second part with $c_2 = \tilde{\lambda}_n \lambda_2^{-1} > 0$.

(iii) We obtain (44) by taking expectation on both sides of (42) and using (ii). □

Acknowledgments. The authors would like to thank the reviewers and the Associated Editors for very useful comments and suggestions. This paper was submitted when both authors were with the Research School of Information Sciences and Engineering, The Australia National University.

REFERENCES

[1] S. Aranda, S. Martinez, and F. Bullo, *On optimal sensor placement and motion coordination for target tracking*, in Proceedings of the IEEE International Conference on Robotics and Automation, Barcelona, Spain, 2005, pp. 4544–4549.

[2] P. Barooah and J. P. Hespanha, *Estimation on graphs from relative measurements: Distributed algorithms and fundamental limits*, IEEE Control Syst. Mag., 27 (2007), pp. 57–74.

[3] R. W. Beard and V. Stepanyan, *Synchronization of information in distributed multiple vehicle coordination control*, in Proceedings of the 42nd IEEE Conference on Decision and Control, Maui, HI, 2003, pp. 2029–2034.

[4] A. Benveniste, M. Métivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, New York, 1990.

[5] P.-A. Bliman and G. Ferrari-Trecate, *Average consensus problems in networks of agents with delayed communications*, in Proceedings of the 44th IEEE Conference on Decision and Control and 2005 European Control Conference, Seville, Spain, 2005, pp. 7066–7071.

[6] V. D. Blondel, J. M. Hendrickx, A. Olshevsky, and J. N. Tsitsiklis, *Convergence in multiagent coordination, consensus, and flocking*, in Proceedings of the 44th IEEE Conference on Decision and Control and 2005 European Control Conference, Seville, Spain, 2005, pp. 2996–3001.

[7] V. Borkar and P. Varaiya, *Asymptotic agreement in distributed estimation*, IEEE Trans. Automat. Control, 27 (1982), pp. 650–655.

[8] A. Böttcher and S. M. Grudsky, *Spectral Properties of Banded Toeplitz Matrices*, SIAM, Philadelphia, 2005.

[9] R. Carli, F. Fagnani, A. Speranzon, and S. Zampieri, *Communication constraints in coordinated consensus problems*, in Proceedings of the American Control Conference, Minneapolis, MN, 2006, pp. 4189–4194.

[10] S. Chatterjee and E. Seneta, *Towards consensus: Some convergence theorems on repeated averaging*, J. Appl. Probab., 14 (1977), pp. 89–97.

[11] H.-F. Chen, *Stochastic Approximation and Its Applications*, Kluwer, Boston, MA, 2002.

[12] Y. S. Chow and T. L. Lai, *Limiting behavior of weighted sums of independent random variables*, Ann. Probab., 1 (1973), pp. 810–824.

[13] Y. S. Chow and H. Teicher, *Probability Theory: Independence, Interchangeability, Martingales*, Springer-Verlag, New York, 1978.

[14] J. Cortes, S. Martinez, and F. Bullo, *Robust rendezvous for mobile autonomous agents via proximity graphs in arbitrary dimensions*, IEEE Trans. Automat. Control, 51 (2006), pp. 1289–1298.

[15] G. A. de Castro and F. Paganini, *Convex synthesis of controllers for consensus*, in Proceedings of the American Control Conference, Boston, MA, 2004, pp. 4933–4938.

[16] P. J. Davis, *Circulant Matrices*, John Wiley, New York, 1979.

[17] J. P. Desai, V. Kumar, and J. P. Ostrowski, *Control of changes in formation for a team of mobile robots*, in Proceedings of the IEEE International Conference on Robotics and Automation, Detroit, MI, 1999, pp. 1556–1561.

[18] M. J. Fischer, *The consensus problem in unreliable distributed systems (a brief survey)*, in Foundations of Computation Theory, M. Karpinsky ed., Lecture Notes in Comput. Sci. 158, Springer-Verlag, New York, 1983, pp. 127–140.

[19] C. Godsil and G. Royle, *Algebraic Graph Theory*, Springer-Verlag, New York, 2001.

[20] M. H. de Groot, *Reaching a consensus*, J. Amer. Statist. Assoc., 69 (1974), pp. 118-121.

[21] Y. Hatano and M. Mesbahi, *Agreement over random networks*, IEEE Trans. Automat. Control, 50 (2005), pp. 1867–1872.

[22] M. Huang and J. H. Manton, *Stochastic approximation for consensus seeking: Mean square and almost sure convergence*, in Proceedings of the 46th IEEE Conference on Decision and Control, New Orleans, LA, 2007, pp. 306–311.

[23] A. Jadbabaie, J. Lin, and A. S. Morse, *Coordination of groups of mobile autonomous agents using nearest neighbor rules*, IEEE Trans. Automat. Control, 48 (2003), pp. 988–1000.

[24] H. J. Kushner and G. G. Yin, *Stochastic Approximation Algorithms and Applications*, Springer-Verlag, New York, 1997.

[25] N. E. Leonard, D. A. Paley, F. Lekien, R. Sepulchre, D. M. Fratantoni, and R. E. Davis, *Collective motion, sensor networks and ocean sampling*, in Proc. IEEE, 95 (2007), pp. 48–74.

[26] Y. Liu and Y. R. Yang, *Reputation propagation and agreement in mobile ad hoc networks*, in Proceedings of the IEEE Wireless Communications and Networking Conference, New Orleans, LA, 2003, pp. 1510–1515.

[27] L. Ljung, *Analysis of recursive stochastic algorithms*, IEEE Trans. Automat. Control, 22 (1977), pp. 551–575.

[28] N. Lynch, *Distributed Algorithms*, Morgan Kaufmann, San Mateo, CA, 1996.

[29] J. A. Marshall and M. E. Broucke, *On invariance of cyclic group symmetries in multiagent formations*, in Proceedings of the 44th IEEE Conference on Decision and Control and 2005 European Control Conference, Seville, Spain, 2005, pp. 746–751.

[30] J. A. Marshall, M. E. Broucke, and B. A. Francis, *Formations of vehicles in cyclic pursuit*, IEEE Trans. Automat. Control, 49 (2004), pp. 1963–1974.

[31] L. Moreau, *Stability of continuous-time distributed consensus algorithms*, in Proceedings of the 43rd IEEE Conference on Decision and Control, Paradise Island, Bahamas, 2004, pp. 3998–4003.

[32] L. Moreau, *Stability of multiagent systems with time-dependent communication links*, IEEE Trans. Automat. Control, 50 (2005), pp. 169–182.

[33] R. Olfati-Saber, *Flocking for multi-agent dynamic systems: Algorithms and theory*, IEEE Trans. Automat. Control, 51 (2006), pp. 401–420.

[34] R. Olfati-Saber and R. M. Murray, *Consensus problems in networks of agents with switching topology and time-delays*, IEEE Trans. Automat. Control, 49 (2004), pp. 1520–1533.

[35] R. Olfati-Saber, J. A. Fax, and R. M. Murray, *Consensus and cooperation in networked multi-agent systems*, Proc. IEEE, 95 (2007), pp. 215–233.

[36] A. V. Oppenheim and R. W. Schafter, *Discrete-Time Signal Processing*, Prentice–Hall, Englewood Cliffs, NJ, 1989.

[37] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications*, Prentice–Hall, Upper Saddle River, NJ, 1996.

[38] W. Ren, R. W. Beard, and E. M. Atkins, *A survey of consensus problems in multi-agent coordination*, in Proceedings of the American Control Conference, Portland, OR, 2005, pp. 1859–1864.

[39] W. Ren, R. W. Beard, and D. B. Kingston, *Multi-agent Kalman consensus with relative uncertainty*, in Proceedings of the American Control Conference, Portland, OR, 2005, pp. 1865–1870.

[40] E. Seneta, *Nonnegative Matrices and Markov Chains*, 2nd ed., Springer-Verlag, New York, 1981.

[41] I. D. Schizas, A. Ribeiro, and G. B. Giannakis, *Consensus in ad hoc WSNs with noisy links—Part* I: *Distributed estimation of deterministic signals*, IEEE Trans. Signal Process., 56 (2008), pp. 350–364.

[42] A. Tahbaz-Salehi and A. Jadbabaie, *On consensus in random networks*, in Proceedings of the 44th Annual Allerton Conference on Communication, Control and Computing, Monticello, IL, 2006, pp. 1315–1321.

[43] H. G. Tanner, G. J. Pappas, and V. Kumar, *Input-to-state stability on formation graphs*, in Proceedings of the 41st IEEE Conference on Decision and Control, Las Vegas, NV, 2002, pp. 2439–2444.

[44] H. G. Tanner, G. J. Pappas, and V. Kumar, *Leader-to-formation stability*, IEEE Trans. Robotics Automat., 20 (2004), pp. 443–455.

[45] H. Teicher, *Almost certain convergence in double arrays*, Z. Wahrsch. Verw. Gebiete, 69 (1985), pp. 331–345.

[46] J. N. Tsitsiklis and M. Athans, *Convergence and asymptotic agreement in distributed decision problems*, IEEE Trans. Automat. Control, 29 (1984), pp. 42–50.

[47] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans, *Distributed asynchronous deterministic and stochastic gradient optimization algorithms*, IEEE Trans. Automat. Control, 31 (1986), pp. 803–812.

[48] P. K. Visscher, *How self-organization evolves*, Nature, 421 (2003), pp. 799–800.

[49] R. L. Winkler, *The consensus of subjective probability distributions*, Management Science, 15 (1968), pp. B61–B75.

[50] L. Xiao and S. Boyd, *Fast linear iterations for distributed averaging*, Systems Control Lett., 53 (2004), pp. 65–78.

[51] L. Xiao, S. Boyd, and S.-J. Kim, *Distributed average consensus with least-mean-square deviation*, J. Parallel Distrib. Comput., 67 (2007), pp. 33–46.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 1, pp. 162–186

© 2009 Society for Industrial and Applied Mathematics

CONTROLLABILITY OF MULTI-AGENT SYSTEMS FROM A GRAPH-THEORETIC PERSPECTIVE*

AMIRREZA RAHMANI[†], MENG JI[‡], MEHRAN MESBAHI[†], AND MAGNUS EGERSTEDT[‡]

Abstract. In this work, we consider the controlled agreement problem for multi-agent networks, where a collection of agents take on leader roles while the remaining agents execute local, consensus-like protocols. Our aim is to identify reflections of graph-theoretic notions on system-theoretic properties of such systems. In particular, we show how the symmetry structure of the network, characterized in terms of its automorphism group, directly relates to the controllability of the corresponding multi-agent system. Moreover, we introduce network equitable partitions as a means by which such controllability characterizations can be extended to the multileader setting.

Key words. multi-agent systems, networked systems, controllability, automorphism group, equitable partitions, agreement dynamics, algebraic graph theory

AMS subject classifications. 93B05, 05C50, 05C25, 34B45, 94C15

DOI. 10.1137/060674909

1. Introduction.

A networked system is a collection of dynamic units that interact over an information exchange network for its operation. Such systems are ubiquitous in diverse areas of science and engineering. Examples include physiological systems and gene networks [12]; large-scale energy systems; and multiple space, air, and land vehicles [1, 2, 20, 27, 37, 38]. There is an active research effort underway in the control and dynamical systems community to study these systems and lay out a foundation for their analysis and synthesis [6, 7, 9, 26]. As a result, over the past few years, a distinct area of research at the intersection of systems theory and graph theory has emerged. An important class of problems that lies at this intersection pertains to the *agreement* or the *consensus problem* [4, 15, 28, 30, 39]. The agreement problem concerns the development of processes by which a group of dynamic units, through local interactions, reach a common value of interest. As such, the agreement protocol is essentially an unforced dynamical system whose trajectory is governed by the interconnection geometry and the initial condition for each unit.

Our goal in this paper is to consider situations where network dynamics can be influenced by external signals and decisions. In particular, we postulate a case involving nodes in the network that do not abide by the agreement protocol; we refer to these agents as *leaders* or *anchors*.[1] The complement of the set of leaders in the network will be referred to as followers (respectively, floating nodes). The presence of these leader nodes generally alters the system behavior. The main topic under consideration in this paper is network controllability when leaders are agents of control. The controllability issue in leader-follower multi-agent systems was introduced in [36]

*Received by the editors October 30, 2006; accepted for publication (in revised form) April 8, 2008; published electronically February 11, 2009. A preliminary version of this work was presented at the 2006 American Control Conference.

http://www.siam.org/journals/sicon/48-1/67490.html

[†]Department of Aeronautics and Astronautics, University of Washington, Seattle, WA 98195 (arahmani@aa.washington.edu, mesbahi@aa.washington.edu). The first and third authors are supported by the National Science Foundation under grant ECS-0501606.

[‡]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 (mengji@ece.gatech.edu, magnus@ece.gatech.edu). The second and fourth authors are supported by the U.S. Army Research Office through grant 99838.

[1] Depending on the context, we could equally consider them as the nonconformist agents.

by Tanner, who provided necessary and sufficient conditions for system controllability in terms of the eigenvectors of the graph Laplacian; we also refer to the related work of Olfati-Saber and Shamma in the context of consensus filters [31]. Subsequently, graph-theoretic characterizations of controllability for leader-follower multi-agent systems were examined by Ji, Muhammad, and Egerstedt [18] and Rahmani and Mesbahi [34]. In the present work, we further explore the ramifications of this graph-theoretic outlook on multi-agent systems controllability. First, we examine the roles of the graph Laplacian eigenvectors and the graph automorphism group for single-leader networks. We then extend these results to multileader setting via equitable partitions of the underlying graph.

This paper begins with the general form of the agreement dynamics over graphs. Next, we introduce transformations that, given the location of the leader nodes, produce the corresponding controlled linear time-invariant system. The study of the controllability for single-leader systems is then pursued via tools from algebraic graph theory. In this venue, we provide a sufficient graphical condition in terms of graph automorphisms for the system's uncontrollability. Furthermore, we introduce network equitable partitions as a means by which such controllability characterizations can be extended to the multileader setting.

2. Notation and preliminaries. In this section we recall some basic notions from graph theory, which is followed by the general setup of the agreement problem for multi-agent networks.

2.1. Graphs and their algebraic representation. Graphs are broadly adopted in the multi-agent literature to encode interactions in networked systems. An undirected graph $\mathcal{G}$ is defined by a set $\mathcal{V}_\mathcal{G} = \{1, \dots, n\}$ of *nodes* and a set $\mathcal{E}_\mathcal{G} \subset \mathcal{V}_\mathcal{G} \times \mathcal{V}_\mathcal{G}$ of *edges*. Two nodes i and j are *neighbors* if $(i, j) \in \mathcal{E}_\mathcal{G}$; the neighboring relation is indicated with $i \sim j$, while $\mathcal{P}(i) = \{j \in \mathcal{V}_\mathcal{G} : j \sim i\}$ collects all neighbors of node i. The *degree* of a node is given by the number of its neighbors; we say that a graph is *regular* if all nodes have the same degree. A *path* $i_0 i_1 \dots i_L$ is a finite sequence of nodes such that $i_{k-1} \sim i_k$, $k = 1, \dots, L$, and a graph $\mathcal{G}$ is *connected* if there is a path between any pair of distinct nodes. A subgraph $\mathcal{G}'$ is said to be *induced* from the original graph $\mathcal{G}$ if it can be obtained by deleting a subset of nodes and edges connecting to those nodes from $\mathcal{G}$.

The *adjacency* matrix of the graph $\mathcal{G}$, $A(\mathcal{G}) \in \mathbb{R}^{n \times n}$, with n denoting the number of nodes in the network, is defined by

$$[\,\mathcal{A}(\mathcal{G})\,]_{ij} := \begin{cases} 1 & \text{if} \quad (i,j) \in \mathcal{E}_\mathcal{G}, \\ 0 & \text{otherwise.} \end{cases}$$

If $\mathcal{G}$ has m edges and is given an arbitrarily orientation, its node-edge *incidence* matrix $\mathcal{B}(\mathcal{G}) \in \mathbb{R}^{n \times m}$ is defined as

$$[\,\mathcal{B}(\mathcal{G})\,]_{kl} := \begin{cases} 1 & \text{if} \quad \text{node } k \text{ is the head of edge } l, \\ -1 & \text{if} \quad \text{node } k \text{ is the tail of edge } l, \\ 0 & \text{otherwise,} \end{cases}$$

where k and l are the indices running over the node and edge sets, respectively.

A matrix that plays a central role in many graph-theoretic treatments of multi-agent systems is the graph *Laplacian*, defined by

$$\mathcal{L}(\mathcal{G}) := \mathcal{B}(\mathcal{G})\,\mathcal{B}(\mathcal{G})^T; \tag{1}$$

thus the graph Laplacian is a (symmetric) positive semidefinite matrix. Let d_i be the degree of node i, and let $\mathcal{D}(\mathcal{G}) := \mathbf{Diag}([d_i]_{i=1}^n)$ be the corresponding diagonal degree matrix. It is easy to verify that $\mathcal{L}(\mathcal{G}) = \mathcal{D}(\mathcal{G}) - \mathcal{A}(\mathcal{G})$ [11]. As the Laplacian is positive semidefinite, its spectrum can be ordered as

$$0 = \lambda_1(\mathcal{L}(\mathcal{G})) \leq \lambda_2(\mathcal{L}(\mathcal{G})) \leq \cdots \leq \lambda_n(\mathcal{L}(\mathcal{G})),$$

with $\lambda_i(\mathcal{L}(\mathcal{G}))$ being the ith ordered eigenvalue of $\mathcal{L}(\mathcal{G})$. It turns out that the multiplicity of the zero eigenvalue of the graph Laplacian is equal to the number of connected components of the graph [14]. In fact the second smallest eigenvalue $\lambda_2(\mathcal{L}(\mathcal{G}))$ provides a judicious measure of the connectivity of $\mathcal{G}$. For more on the related matrix-theoretic and algebraic approaches to graph theory, we refer the reader to [5, 14, 24].

2.2. Agreement dynamics. Given a multi-agent system with n agents, we can model the network by a graph $\mathcal{G}$ where nodes represent agents and edges are inter-agent information exchange links.[2] Let $x_i(t) \in \mathbb{R}^d$ denote the state of node i at time t, whose dynamics is described by the single integrator

$$\dot{x}_i(t) = u_i(t), \qquad i = 1, \ldots, n,$$

with $u_i(t)$ being node i's control input. Next, we allow agent i to have access to the relative state information with respect to its neighbors and use it to compute its control. Hence, interagent coupling is realized through $u_i(t)$. For example, one can let

$$u_i(t) = -\sum_{i\sim j}(x_i(t) - x_j(t)). \tag{2}$$

The localized rule in (2) happens to lead to the solution of the rendezvous problem, which has attracted considerable attention in the literature [8, 17, 22]. Some other important networked system problems, e.g., formation control [3, 10, 13], consensus or agreement [25, 29, 30], and flocking [32, 35], share the same distributive flavor as the rendezvous problem.

The single integrator dynamics in conjunction with (2) can be represented as the Laplacian dynamics of the form

$$\dot{x}(t) = -\mathbb{L}(\mathcal{G})x(t), \tag{3}$$

where $x(t) = [x(t)_1^T,\ x(t)_2^T, \ldots, x(t)_n^T]^T$ denotes the aggregated state vector of the multi-agent system, $\mathbb{L}(\mathcal{G}) := \mathcal{L}(\mathcal{G}) \otimes I_d$, with I_d denoting the d-dimensional identity matrix, and $\otimes$ is the matrix Kronecker product [16]. In fact, if the dynamics of the agent's state is decoupled along each dimension, the behavior of the multi-agent system can be investigated one dimension at a time. Although our results can directly be extended to the case of (3), in what follows we will focus on the system

$$\dot{x}(t) = -\mathcal{L}(\mathcal{G})x(t), \tag{4}$$

capturing the multi-agent dynamics with individual agent states evolving in $\mathbb{R}$.

[2] Throughout this paper we assume that the network is static. As such, the movements of the agents will not cause edges to appear or disappear in the network.

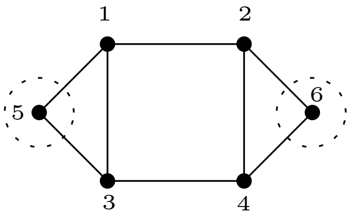


FIG. 1. *A leader-follower network with* $\mathcal{V}_f = \{1, 2, 3, 4\}$ *and* $\mathcal{V}_l = \{5, 6\}$.

3. Controlled agreement. We now endow leadership roles to a subset of agents in the Laplacian dynamics (4); the other agents in the network, the followers, continue to abide by the agreement protocol. In this paper, we use subscripts l and f to denote affiliations with leaders and followers, respectively. For example, a *follower graph* $\mathcal{G}_f$ is the subgraph induced by the follower node set $\mathcal{V}_f \subset \mathcal{V}_{\mathcal{G}}$. Leadership designations induce a partition of incidence matrix $\mathcal{B}(\mathcal{G})$ as

$$\mathcal{B}(\mathcal{G}) = \begin{bmatrix} \mathcal{B}_f(\mathcal{G}) \\ \mathcal{B}_l(\mathcal{G}) \end{bmatrix}, \tag{5}$$

where $\mathcal{B}_f(\mathcal{G}) \in \mathbb{R}^{n_f \times m}$, and $\mathcal{B}_l(\mathcal{G}) \in \mathbb{R}^{n_l \times m}$. Here n_f and n_l are the cardinalities of the follower group and the leader group, respectively, and m is the number of edges. The underlying assumption of this partition, without loss of generality, is that leaders are indexed last in the original graph $\mathcal{G}$. As a result of (1) and (5), the graph Laplacian $\mathcal{L}(\mathcal{G})$ is given by

$$\mathcal{L}(\mathcal{G}) = \begin{bmatrix} \mathcal{L}_f(\mathcal{G}) & l_{fl}(\mathcal{G}) \\ l_{fl}(\mathcal{G})^T & \mathcal{L}_l(\mathcal{G}) \end{bmatrix}, \tag{6}$$

where

$$\mathcal{L}_f(\mathcal{G}) = \mathcal{B}_f\mathcal{B}_f^T, \quad \mathcal{L}_l(\mathcal{G}) = \mathcal{B}_l\mathcal{B}_l^T, \quad \text{and } l_{fl}(\mathcal{G}) = \mathcal{B}_f\mathcal{B}_l^T.$$

Here we omitted the dependency of $\mathcal{B}, \mathcal{B}_f$, and $\mathcal{B}_l$ on $\mathcal{G}$, which we will continue to do whenever this dependency is clear from the context. As an example, Figure 1 shows a leader-follower network with $\mathcal{V}_l = \{5, 6\}$ and $\mathcal{V}_f = \{1, 2, 3, 4\}$. This gives

$$\mathcal{B}_f = \begin{bmatrix} 1 & 0 & 0 & -1 & 0 & 1 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & -1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 1 & -1 & 0 & 0 & 0 \end{bmatrix}, \quad \mathcal{B}_l = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix},$$

and

$$\mathcal{L}_f(\mathcal{G}) = \begin{bmatrix} 3 & -1 & 0 & -1 \\ -1 & 3 & -1 & 0 \\ 0 & -1 & 3 & -1 \\ -1 & 0 & -1 & 3 \end{bmatrix}, \qquad l_{fl}(\mathcal{G}) = \begin{bmatrix} -1 & 0 \\ 0 & -1 \\ 0 & -1 \\ -1 & 0 \end{bmatrix}.$$

The control system we now consider is the controlled agreement dynamics (or leader-follower system), where followers evolve through the Laplacian-based dynamics

$$\dot{x}_f(t) = -\mathcal{L}_f\, x_f(t) - l_{fl}\, u(t), \tag{7}$$

where u denotes the exogenous control signal dictated by the leaders' states.

DEFINITION 3.1. *Let l be a leader node in $\mathcal{G}$, i.e., $l \in \mathcal{V}_l(\mathcal{G})$. The indicator vector with respect to l,*

$$\delta_l : \mathcal{V}_f \to \{0,1\}^{n_f},$$

is such that

$$\delta_l(i) := \begin{cases} 1 & \textit{if} \quad i \sim l, \\ 0 & \textit{otherwise.} \end{cases}$$

We note that each column of $-l_{fl}$ is an indicator vector, i.e., $l_{fl} = [-\delta_{n_f+1}, \ldots, -\delta_n]$.

Let d_{il}, with $i \in \mathcal{V}_f$, denote the number of leaders adjacent to follower i, and define the follower-leader degree matrix

$$\mathcal{D}_{fl}(\mathcal{G}) := \mathbf{Diag}([d_{il}]_{i=1}^{n_f}), \tag{8}$$

which leads to the relationship

$$\mathcal{L}_f(\mathcal{G}) = \mathcal{L}(\mathcal{G}_f) + \mathcal{D}_{fl}(\mathcal{G}), \tag{9}$$

where $\mathcal{L}(\mathcal{G}_f)$ is the Laplacian matrix of the follower graph $\mathcal{G}_f$.

Remark 3.2. We should emphasize the difference between $\mathcal{L}_f(\mathcal{G})$ and $\mathcal{L}(\mathcal{G}_f)$. The matrix $\mathcal{L}_f(\mathcal{G})$ is the principle diagonal submatrix of the original Laplacian matrix $\mathcal{L}(\mathcal{G})$ related to the followers, while $\mathcal{L}(\mathcal{G}_f)$ is the Laplacian matrix of the subgraph $\mathcal{G}_f$ induced by the followers. For simplicity, we will write $\mathcal{L}_f$ and l_{fl} to represent $\mathcal{L}_f(\mathcal{G})$ and $l_{fl}(\mathcal{G})$, respectively, when their dependency on $\mathcal{G}$ is clear from the context.

Since the row sum of the Laplacian matrix is zero, the sum of the ith row of $\mathcal{L}_f(\mathcal{G})$ and that of $-l_{fl}(\mathcal{G})$ are both equal to d_{il}, i.e.,

$$\mathcal{L}_f(\mathcal{G})\,\mathbf{1}_{n_f} = \mathcal{D}_{fl}(\mathcal{G})\,\mathbf{1}_{n_f} = -l_{fl}(\mathcal{G})\,\mathbf{1}_{n_l}, \tag{10}$$

where $\mathbf{1}$ is a vector with ones at each component.

If there is only one leader in the network, then according to the indexing convention, $\mathcal{V}_l = \{n\}$. In this case, we have $l_{fl}(\mathcal{G}) = -\delta_n$ and $\mathcal{D}_{fl}(\mathcal{G}) = \mathbf{Diag}(\delta_n)$. For instance, the indicator vector for the node set $\mathcal{V}_f = \{1,2,3\}$ in the graph shown in Figure 2 with respect to the leader $\{4\}$ is $\delta_4 = [\,1,\,1,\,0\,]^T$.

PROPOSITION 3.3. *If a single node is chosen to be the leader, the original Laplacian $\mathcal{L}(\mathcal{G})$ is related to the Laplacian of the follower graph $\mathcal{L}(\mathcal{G}_f)$ via*

$$\mathcal{L}(\mathcal{G}) = \begin{bmatrix} \mathcal{L}(\mathcal{G}_f) + \mathcal{D}_{fl}(\mathcal{G}) & -\delta_n \\ -\delta_n^T & d_n \end{bmatrix}, \tag{11}$$

where d_n denotes the degree of agent n.

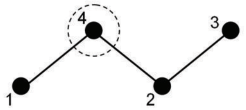


FIG. 2. *Path graph with node "4" being the leader.*

Another way to construct the system matrices $\mathcal{L}_f(\mathcal{G})$ and $l_{fl}(\mathcal{G})$ is from the Laplacian of the original graph via

$$\mathcal{L}_f = P_f^T\,\mathcal{L}(\mathcal{G})\,P_f \;\text{ and }\; l_{fl} = P_f^T\,\mathcal{L}(\mathcal{G})\,T_{fl}, \tag{12}$$

where $P_f \in \mathbb{R}^{n\times n_f}$ is constructed by eliminating the columns of the $n \times n$ identity matrix that correspond to the leaders, and $T_{fl} \in \mathbb{R}^{n\times n_l}$ is formed by grouping these eliminated columns into a new matrix. For example, in Figure 1 these matrices assume the form

$$P_f = \begin{bmatrix} I_4 \\ \mathbf{0}_{2\times 4} \end{bmatrix} \quad \text{and} \quad T_{fl} = \begin{bmatrix} \mathbf{0}_{4\times 2} \\ I_2 \end{bmatrix}.$$

PROPOSITION 3.4. *If a single node is chosen to be the leader, one has*

$$T_{fl} = (I_n - \tilde{P})\mathbf{1}_n \quad \textit{and} \quad l_{fl} = -\mathcal{L}_f \mathbf{1}_{n_f}$$

in (12)*, where* $\tilde{P} = [P_f \quad \mathbf{0}_{n\times n_l}]$ *is the* $n \times n$ *square matrix obtained by expanding* P_f *with zero block of proper dimensions.*

Proof. The first equality directly follows from the definition of P_f and T_{fl}. Without loss of generality, assume that the last node is the leader; then $[\,P_f \quad T_{fl}\,] = I_n$. Multiplying both sides by $\mathbf{1}_n$ and noting that $\tilde{P}\,\mathbf{1}_n = P_f \mathbf{1}_{n_f}$, one has $T_{fl} = (I_n - \tilde{P})\mathbf{1}_n$. Moreover,

$$l_{fl} = P_f^T \mathcal{L}(\mathcal{G})\{(I - \tilde{P})\mathbf{1}_n\} = P_f^T \mathcal{L}(\mathcal{G})\mathbf{1}_n - P_f^T \mathcal{L}(\mathcal{G}) P_f \mathbf{1}_{n_f}.$$

The first term on the right-hand side of the equality is zero, as $\mathbf{1}$ belongs to the null space of $\mathcal{L}(\mathcal{G})$; the second term, on the other hand, is simply $\mathcal{L}_f \mathbf{1}$. □

Alternatively, for the case when the exogenous signal is constant, the dynamics (7) can be rewritten as

$$\begin{bmatrix} \dot{x}_f(t) \\ \dot{u}(t) \end{bmatrix} = -\begin{bmatrix} \mathcal{L}_f & l_{fl} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} x_f(t) \\ u(t) \end{bmatrix}. \tag{13}$$

This corresponds to zeroing-out the rows of the original graph Laplacian associated with the leader. Zeroing-out a row of a matrix can be accomplished via a reduced identity matrix Q_r, with zeros at the diagonal elements that correspond to the leaders, with all other diagonal elements being kept as one. In this case,

$$\begin{bmatrix} \mathcal{L}_f & l_{fl} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = Q_r \mathcal{L}(\mathcal{G}), \tag{14}$$

where

$$Q_r = \begin{bmatrix} I_{n_f} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

and all the zero matrices are of proper dimensions.

4. Reachability. First, we examine whether we can steer the system (7) into the agreement subspace, $\mathbf{span}\{\mathbf{1}\}$, when the exogenous signal is constant, i.e., $x_i = c$, for all $i \in \mathcal{V}_l$ and $c \in \mathbb{R}$ is a constant. As shown in (14), in this case the controlled agreement can be represented as

$$\dot{x}(t) = -Q_r \mathcal{L}(\mathcal{G})\, x(t) = -\mathcal{L}_r(\mathcal{G})\, x(t), \tag{15}$$

where Q_r is the reduced identity matrix and $\mathcal{L}_r(\mathcal{G}) := Q_r \mathcal{L}(\mathcal{G})$ is the reduced Laplacian matrix. Let us now examine the convergence properties of (15) with respect to

span$\{\mathbf{1}\}$. Define $\zeta(t)$ as the projection of the followers' state $x_f(t)$ onto the subspace orthogonal to the agreement subspace **span**$\{\mathbf{1}\}$. This subspace is denoted by $\mathbf{1}^{\perp}$; in [30] it is referred to as the *disagreement* subspace. One can then model the disagreement dynamics as

$$\dot{\zeta}(t) = -\mathcal{L}_r(\mathcal{G})\,\zeta(t). \tag{16}$$

Choosing a standard quadratic Lyapunov function for (16),

$$V(\zeta(t)) = \frac{1}{2}\,\zeta(t)^T\zeta(t),$$

reveals that its time rate of change assumes the form

$$\dot{V}(\zeta(t)) = -\zeta(t)^T\,\overline{\mathcal{L}}_r(\mathcal{G})\,\zeta(t),$$

where $\overline{\mathcal{L}}_r(\mathcal{G}) = (1/2)\,[\,\mathcal{L}_r(\mathcal{G}) + \mathcal{L}_r(\mathcal{G})^T\,]$.

PROPOSITION 4.1. *The agreement subspace is reachable for the controlled agreement protocol* (7).

Proof. Since $\dot{V}(\zeta) < 0$ for all $\zeta \neq 0$ and $Q_r\mathcal{L}(\mathcal{G})\,\mathbf{1} = 0$, for any leader nodes, the agreement subspace remains a globally attractive subspace of (15). □

PROPOSITION 4.2. *In the case of one leader, the matrix $\mathcal{L}_r(\mathcal{G})$ has a real spectrum and the same inertia as $\mathcal{L}(\mathcal{G})$.*

Proof. Let $E = \mathbf{1}\mathbf{1}^T$ denote the matrix of all ones. Since $E\mathcal{L}(\mathcal{G}) = 0$ and $Q_r\mathcal{L}(\mathcal{G}) = \mathcal{L}_r(\mathcal{G})$, $(Q_r + E)\,\mathcal{L}(\mathcal{G}) = \mathcal{L}_r(\mathcal{G})$. Hence $\mathcal{L}_r(\mathcal{G})$ is a product of a positive definite matrix, namely Q_r+E, and the symmetric matrix $\mathcal{L}(\mathcal{G})$. By Theorem 7.6.3 of [16], $\mathcal{L}_r(\mathcal{G})$ is diagonalizable and has a real spectrum. In fact, it has the same inertia as $\mathcal{L}(\mathcal{G})$. □

5. Controllability analysis of single-leader networks.

In this section, we investigate the controllability properties of single-leader networks. Following our previously mentioned indexing convention, the index of the leader is assumed to be n. For notational convenience in this section, we will equate x_f with x and x_l with u. Moreover, we identify matrices A and B with $-\mathcal{L}_f$ and $-l_{fl}$, respectively. Thus, the system (7) is specified by

$$\dot{x}(t) = Ax(t) + Bu(t). \tag{17}$$

The controllability of the controlled agreement (17) can be investigated using the Popov–Hautus–Belevitch (PHB) test [19, 33]. Specifically, (17) is uncontrollable if and only if there exists a left eigenvector ν of A, i.e., $\nu^T A = \lambda\nu^T$ for some λ, such that

$$\nu^T B = 0.$$

Since A is symmetric, its left and right eigenvectors are the same. Hence, the necessary and sufficient condition for controllability of (17) is that none of the eigenvectors of A should be simultaneously orthogonal to all columns of B. Additionally, in order to investigate the controllability of (17), one can form the controllability matrix as

$$\mathcal{C} = [\,B \;\; AB \;\; \cdots \;\; A^{n_f-1}B\,]. \tag{18}$$

As A is symmetric, it can be written in the form $U\Lambda U^T$, where Λ is the diagonal matrix of eigenvalues of A; U, on the other hand, is the unitary matrix comprised of

A's pairwise orthogonal unit eigenvectors. Since $B = UU^TB$, by factoring the matrix U from the left in (18), the controllability matrix assumes the form

$$\mathcal{C} = U\ [U^TB\ \ \Lambda U^TB\ \dots\ \Lambda^{n_f-1}U^TB]. \tag{19}$$

In this case, U is full rank and its presence does not alter the rank of the matrix product in (19). If one of the columns of U is perpendicular to all the columns of B, then $\mathcal{C}$ will have a row equal to zero, and hence the matrix $\mathcal{C}$ is rank deficient [36]. On the other hand, in the case of one leader, if any two eigenvalues of A are equal, then $\mathcal{C}$ will have two linear dependent rows, and again, the controllability matrix becomes rank deficient. Assume that ν_1 and ν_2 are two eigenvectors corresponding to the same eigenvalue and that none of them is orthogonal to B. Then $\nu = \nu_1 + c\nu_2$ is also an eigenvector of A for that eigenvalue. This will then allow us to choose $c = -\nu_1^TB/\nu_2^TB$, which renders $\nu^TB = 0$. In other words, we are able to find an eigenvector that is orthogonal to B. Hence, we arrive at the following observation.

Proposition 5.1. *Consider a leader-follower network whose evolution is described by* (17). *This system is controllable if and only if none of the eigenvectors of* A *is (simultaneously) orthogonal to (all columns of)* B. *Moreover, if* A *does not have distinct eigenvalues, then* (17) *is not controllable.*

Proposition 5.1 is also valid for the case with more than one leader and implies that in any finite time interval, the floating dynamic units can be independently steered from their initial states to an arbitrary final one based on local interactions with their neighbors. This controllability results is of course valid when the states of the leader nodes are assumed to be unconstrained.

Corollary 5.2. *The networked system* (17) *with a single leader is controllable if and only if none of the eigenvectors of* A *is orthogonal to* $\mathbf{1}$.

Proof. As shown in Proposition 3.4, the elements of B correspond to row-sums of A, i.e., $B = -A\mathbf{1}$. Thus, $\nu^TB = -\nu^TA\mathbf{1} = -\lambda\,(\nu^T\mathbf{1})$. By Proposition 4.2 one has $\lambda \neq 0$. Thereby, $\nu^TB = 0$ if and only if $\mathbf{1}^T\nu = 0$. □

Proposition 5.3. *If the networked system* (17) *is uncontrollable, there exists an eigenvector* ν *of* A *such that* $\sum_{i\sim n}\nu(i) = 0$.

Proof. Using Corollary 5.2, when the system is uncontrollable, there exists an eigenvector of A that is orthogonal to $\mathbf{1}$. As $A\nu = \lambda\nu$, we deduce that $\mathbf{1}^T(\,A\,\nu\,) = 0$. Moreover, using Proposition 3.3, we obtain

$$\nu^T\,\{\,\mathcal{L}(\mathcal{G}_f) + \mathcal{D}_{fl}(\mathcal{G})\,\}\;\mathbf{1} = \,0.$$

But $\mathcal{L}(\mathcal{G}_f)\mathbf{1} = 0$, and thereby

$$\nu^T\,\mathcal{D}_{fl}(\mathcal{G})\mathbf{1} = \nu^T\delta_n = 0,$$

which implies that $\sum_{i\sim n}\nu(i) = 0$. □

Proposition 5.4. *Suppose that the leader-follower system* (17) *is uncontrollable. Then there exists an eigenvector of* $\mathcal{L}(\mathcal{G})$ *that has a zero component on the index that corresponds to the leader.*

Proof. Let ν be an eigenvector of A that is orthogonal to $\mathbf{1}$ (by Corollary 5.2, such an eigenvector exists). Attach a zero to ν; using the partitioning (11), we then have

$$\mathcal{L}(\mathcal{G})\begin{bmatrix}\nu\\0\end{bmatrix} = \begin{bmatrix}A & -\delta_n\\ -\delta_n^T & d_n\end{bmatrix}\begin{bmatrix}\nu\\0\end{bmatrix} = \begin{bmatrix}\lambda\nu\\ -\delta_n^T\nu\end{bmatrix},$$

where δ_n is the indicator vector of the leader's neighbors. From Proposition 5.3 we know that $\delta_n^T \nu = 0$. Thus,

$$\mathcal{L}(\mathcal{G}) \begin{bmatrix} \nu \\ 0 \end{bmatrix} = \lambda \begin{bmatrix} \nu \\ 0 \end{bmatrix}.$$

In the other words, $\mathcal{L}(\mathcal{G})$ has an eigenvector with a zero on the index that corresponds to the leader. □

A direct consequence of Proposition 5.4 is the following.

COROLLARY 5.5. *Suppose that none of the eigenvectors of* $\mathcal{L}(\mathcal{G})$ *has a zero component. Then the leader-follower system* (7) *is controllable for any choice of the leader.*

5.1. Controllability and graph symmetry. The controllability of the interconnected system depends not only on the geometry of the interunit information exchange but also on the position of the leader with respect to the graph topology. In this section, we examine the controllability of the system in terms of graph-theoretic properties of the network. In particular, we will show that there is an intricate relation between the controllability of (17) and the symmetry structure of the graph, as captured by its automorphism group. We first need to introduce a few useful constructs.

DEFINITION 5.6. *A permutation matrix is a* $\{0, 1\}$*-matrix with a single nonzero element in each row and column.*

DEFINITION 5.7. *The system* (17) *is anchor symmetric with respect to anchor* a *if there exists a nonidentity permutation* J *such that*

$$JA = AJ, \tag{20}$$

where $A = -\mathcal{L}_f = -P_f^T \, \mathcal{L}(\mathcal{G}) \, P_f$ *is constructed as in* (12)*. We call the system asymmetric if it does not admit such a permutation for any anchor.*

As an example, the graph represented in Figure 3(a) is leader symmetric with respect to $\{6\}$ but asymmetric with respect to any other leader node set. On the other hand, the graph of Figure 3(b) is leader symmetric with respect to a single leader located at every node. The utility of the notion of leader symmetry is now established through its relevance to the system-theoretic concept of controllability.

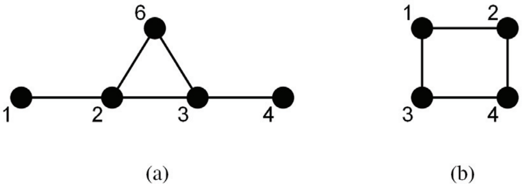


FIG. 3. *Interconnected topologies that are leader symmetric:* (a) *only with respect to node* $\{6\}$*;* (b) *with respect to a leader at any node.*

PROPOSITION 5.8. *The system* (17) *is uncontrollable if it is leader symmetric.*

Proof. If the system is leader symmetric, then there is a nonidentity permutation J such that

$$JA = AJ. \tag{21}$$

Recall that, by Proposition 5.1, if the eigenvalues of A are not distinct, then (17) is not controllable. We thus consider the case where all eigenvalues λ are distinct and satisfy

$A\nu = \lambda\nu$; thereby, for all eigenvalue/eigenvector pairs (λ, ν) one has $JA\nu = J(\lambda\nu)$. Using (21), however, we see that $A\,(J\nu) = \lambda\,(J\nu)$, and $J\nu$ is also an eigenvector of A corresponding to the eigenvalue λ. Given that λ is distinct and A admits a set of orthonormal eigenvectors, we conclude that for one such eigenvector ν, $\nu - J\nu$ is also an eigenvector of A. Moreover, $J\,B = J^T B = B$, as the elements of B correspond to the row-sums of the matrix A, i.e., $B = -A\,\mathbf{1}$. Thereby,

$$(\nu - J\nu)^T B = \nu^T B - \nu^T J^T\,B = \nu^T B - \nu^T B = 0. \tag{22}$$

This, on the other hand, translates into having one of the eigenvectors of A, namely $\nu - J\nu$, be orthogonal to B. Proposition 5.1 now implies that the system (17) is uncontrollable. □

Proposition 5.8 states that leader symmetry is a sufficient condition for uncontrollability of the system. It is instructive to examine whether leader asymmetry leads to a controllable system.

PROPOSITION 5.9. *Leader symmetry is not a necessary condition for system uncontrollability.*

Proof. In Figure 4, the subgraph shown by solid lines, $\mathcal{G}_f$, is the smallest asymmetric graph [21], in the sense that it does not admit any nonidentity automorphism. Let us augment this graph with the node "a" and connect it to all vertices of $\mathcal{G}_f$. Constructing the corresponding system matrix A (i.e., setting it equal to $-\mathcal{L}_f(\mathcal{G})$), we have

$$-A = \mathcal{L}(\mathcal{G}_f) + \mathcal{D}_{fl}(\mathcal{G}) = \mathcal{L}(\mathcal{G}_f) + I,$$

where I is the identity matrix of proper dimensions. Consequently, A has the same set of eigenvectors as $\mathcal{L}(\mathcal{G}_f)$. Since $\mathcal{L}(\mathcal{G}_f)$ has an eigenvector orthogonal to $\mathbf{1}$, A also has an eigenvector that is orthogonal to $\mathbf{1}$. Hence, the leader-follower system is not controllable. Yet, the system is not symmetric with respect to a; more on this will appear in section 5.2. □

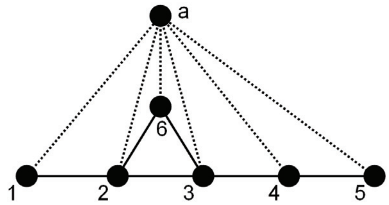


FIG. 4. *Asymmetric information topology with respect to the leader* $\{a\}$. *The subgraph shown by solid lines is the smallest asymmetric graph.*

It is intuitive that a highly connected leader will result in faster convergence to the agreement subspace. However, a highly connected leader also increases the chances that a symmetric graph, with respect to leader, emerges. A limiting case for this latter scenario is the complete graph. In such a graph, $n - 1$ leaders are needed to make the corresponding controlled system controllable. This requirement is of course not generally desirable, as it means that the leader group includes all nodes except for one node! The complete graph is "the worse" case configuration as far as its controllability properties. Generally at most $n - 1$ leaders are needed to make any information exchange network controllable. In the meantime, a path graph with a leader at one end is controllable. Thus it is possible to make a complete graph controllable by keeping the links on the longest path between a leader and all other nodes and deleting the unnecessary information exchange links to break its inherent symmetry. This procedure is not always feasible; for example, a star graph is not amenable to such graphical alterations.

5.2. Leader symmetry and graph automorphism. In section 5.1 we discussed the relationship between leader symmetry and controllability. In this section we will further explore the notion of leader symmetry with respect to graph automorphisms.

Definition 5.10. *An automorphism of $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a permutation ψ of its node set such that*

$$(\psi(i)\,, \psi(j)) \in \mathcal{E}_{\mathcal{G}} \quad \Longleftrightarrow \quad (i, j) \in \mathcal{E}_{\mathcal{G}}.$$

The set of all automorphisms of $\mathcal{G}$, equipped with the composition operator, constitutes the automorphism group of $\mathcal{G}$; note that this is a “finite” group. It is clear that the degree of a node remains unchanged under the action of the automorphism group; i.e., if ψ is an automorphism of $\mathcal{G}$, then $d_v = d_{\psi(v)}$ for all $v \in \mathcal{V}_{\mathcal{G}}$.

Proposition 5.11 (see [5]). *Let $\mathcal{A}(\mathcal{G})$ be the adjacency matrix of the graph $\mathcal{G}$ and ψ a permutation on its node set $\mathcal{V}$. Associate with this permutation the permutation matrix Ψ such that*

$$\Psi_{ij} := \begin{cases} 1 & \textit{if} \quad \psi(i) = j, \\ 0 & \textit{otherwise}. \end{cases}$$

Then ψ is an automorphism of $\mathcal{G}$ if and only if

$$\Psi\,\mathcal{A}(\mathcal{G}) \;=\; \mathcal{A}(\mathcal{G})\,\Psi.$$

In this case, the least positive integer z for which $\Psi^z = I$ is called the order of the automorphism.

Recall that from Definition 5.7 leader symmetry for (17) corresponds to having

$$JA = AJ,$$

where J is a nonidentity permutation. From Proposition 3.3, however,

$$A = -(\mathcal{L}(\mathcal{G}_f) + \mathcal{D}_{fl}(\mathcal{G})).$$

Thus using the identity $\mathcal{L}(\mathcal{G}_f) = \mathcal{D}(\mathcal{G}_f) - \mathcal{A}(\mathcal{G}_f)$, one has

$$J\,\{\mathcal{D}(\mathcal{G}_f) - \mathcal{A}(\mathcal{G}_f) + \mathcal{D}_{fl}(\mathcal{G})\} = \{\mathcal{D}(\mathcal{G}_f) - \mathcal{A}(\mathcal{G}_f) + \mathcal{D}_{fl}(\mathcal{G})\}\; J. \tag{23}$$

Pre- and postmultiplication of (a permutation matrix) J does not change the structure of diagonal matrices. Also, all diagonal elements of $\mathcal{A}(\mathcal{G})$ are zero. We can thereby rewrite (23) as two separate conditions,

$$J\mathcal{D}_f(\mathcal{G}) = \mathcal{D}_f(\mathcal{G})\,J \quad \text{and} \quad J\mathcal{A}(\mathcal{G}_f) = \mathcal{A}(\mathcal{G}_f)J, \tag{24}$$

with $\mathcal{D}_f(\mathcal{G}) := \mathcal{D}(\mathcal{G}_f) + \mathcal{D}_{fl}(\mathcal{G})$. The second equality in (24) states that sought after J in (20) is in fact an automorphism of $\mathcal{G}_f$.

Proposition 5.12. *Let Ψ be the permutation matrix associated with ψ. Then $\Psi\,\mathcal{D}_f(\mathcal{G}) = \mathcal{D}_f(\mathcal{G})\,\Psi$ if and only if*

$$d_i + \delta_n(i) = d_{\psi(i)} + \delta_n(\psi(i)).$$

In the case where ψ is an automorphism of $\mathcal{G}_f$, this condition simplifies to

$$\delta_n(i) = \delta_n(\psi(i)).$$

Proof. Using the properties of permutation matrices, one has

$$[\Psi\mathcal{D}_f(\mathcal{G})]_{ik} = \sum_t \Psi_{it}\mathcal{D}_{tk} = \begin{cases} d_k + \delta_n(k) & \text{if} \quad i \to k, \\ 0 & \text{otherwise} \end{cases}$$

and

$$[\mathcal{D}_f(\mathcal{G})\Psi]_{ik} = \sum_t \mathcal{D}_{it}\,\Psi_{tk} = \begin{cases} d_i + \delta_n(i) & \text{if} \quad i \to k, \\ 0 & \text{otherwise.} \end{cases}$$

For these matrices to be equal elementwise, one needs to have $d_i + \delta_n(i) = d_k + \delta_n(k)$ when $\psi(i) = k$. The second statement in the proposition follows from the fact that the degree of a node remains invariant under the action of the automorphism group. □

The next two results follow immediately from the above discussion.

Proposition 5.13. *The interconnected system* (17) *is leader symmetric if and only if there is a nonidentity automorphism for $\mathcal{G}_f$ such that the indicator function remains invariant under its action.*

Corollary 5.14. *The interconnected system* (17) *is leader asymmetric if the automorphism group of the floating (or follower) subgraph contains only the trivial (identity) permutation.*

5.3. Controllability of special graphs. In this section we investigate the controllability of ring and path graphs.

Proposition 5.15. *A ring graph, with only one leader, is never controllable.*

Proof. With only one leader in the ring graph, the follower graph $\mathcal{G}_f$ becomes the path graph with one nontrivial automorphism, i.e., its mirror image. Without loss of generality, choose the first node as the leader and index the remaining follower nodes by a clockwise traversing of the ring. Then the permutation $i \to n - i + 2$ for $i = 2, \dots, n$ is an automorphism of $\mathcal{G}_f$. In the meantime, the leader "1" is connected to both node 2 and node n; hence $\delta_n = [1, 0, \dots, 0, 1]^T$ remains invariant under the permutation. Using Proposition 5.13, we conclude that the corresponding system (17) is leader symmetric and thus uncontrollable. □

Proposition 5.16. *A path graph is controllable for all choices of the leader node if and only if it is of even order.*

Proof. Suppose that the path graph is of odd order; then choose the middle node $\frac{n+1}{2}$ as the leader. Note that $\psi(k) = n - k + 1$ is an automorphism for the floating subgraph. Moreover, the leader is connected to nodes $\frac{n+1}{2} - 1$ and $\frac{n+1}{2} + 1$, and $\psi(\frac{n+1}{2} - 1) = \frac{n+1}{2} + 1$. Thus

$$\delta_n = [0, \dots, 0, 1, 1, 0, \dots, 0]^T$$

remains invariant under the permutation ψ and the system is uncontrollable. The converse statement follows analogously. □

Hence although in general leader symmetry is a sufficient—yet not necessary—condition for uncontrollability of (17), it is necessary and sufficient for uncontrollability of the path graph.

Corollary 5.17. *A path graph with a single leader is controllable if and only if it is leader asymmetric.*

6. Rate of convergence. In previous sections, we discussed controllability properties of controlled agreement dynamics in terms of the symmetry structure of the network. When the resulting system is controllable, the nodes can reach agreement arbitrarily fast.

Proposition 6.1. *A controllable agreement dynamics* (17) *can reach the agreement subspace arbitrarily fast.*

Proof. The (invertible) controllability Grammian for (17) is defined as

$$W_a(t_0, t_f) = \int_{t_0}^{t_f} e^{sA} B B^T e^{sA^T}\, ds. \tag{25}$$

For any $t_f > t_0$, the leader can then transmit the signal

$$u(t) = B^T e^{A^T(t_f - t_0)} W_a(t_0, t_f)^{-1} \left(x_f - e^{A(t_f - t_0)} x_0 \right) \tag{26}$$

to its neighbors; in (26) x_0 and x_f are the initial and final states for the follower nodes, and t_0 and t_f are prespecified initial and final maneuver times. □

Next let us examine the convergence properties of the leader-follower network with a leader that transmits a constant signal (15). In this venue, define the quantity

$$\mu_2(\mathcal{L}_r(\mathcal{G})) := \min_{\substack{\zeta \neq 0 \\ \zeta \perp \mathbf{1}}} \frac{\zeta^T \, \overline{\mathcal{L}}_r(\mathcal{G}) \, \zeta}{\zeta^T \zeta}. \tag{27}$$

Proposition 6.2. *The rate of convergence of the disagreement dynamics* (16) *is bounded by $\mu_2(\mathcal{L}_r(\mathcal{G}))$ and $\lambda_2(\mathcal{L}(\mathcal{G}))$, when the leader transmits a constant signal.*

Proof. Using the variational characterization of the second smallest eigenvalue of the graph Laplacian [14, 16], we have

$$\begin{aligned}
\lambda_2(\mathcal{L}(\mathcal{G})) &= \min_{\substack{\zeta \neq 0 \\ \zeta \perp \mathbf{1}}} \frac{\zeta^T \mathcal{L}(\mathcal{G}) \zeta}{\zeta^T \zeta} \leq \min_{\substack{\zeta \neq 0 \\ \zeta \perp \mathbf{1} \\ \zeta = Q\beta}} \frac{\zeta^T \mathcal{L}(\mathcal{G}) \zeta}{\zeta^T \zeta} \\
&= \min_{\substack{Q\beta \neq 0 \\ Q\beta \perp \mathbf{1}}} \frac{\beta^T Q \mathcal{L}(\mathcal{G}) Q \beta}{\beta^T Q \beta} \\
&= \min_{\substack{Q\beta \neq 0 \\ Q\beta \perp \mathbf{1}}} \frac{\beta^T Q \left\{ \frac{1}{2}(Q\mathcal{L}(\mathcal{G}) + \mathcal{L}(\mathcal{G})Q) \right\} Q\beta}{\beta^T Q \beta} \\
&= \min_{\substack{Q\beta \neq 0 \\ Q\beta \perp \mathbf{1}}} \frac{\beta^T Q \left(\frac{1}{2}(\mathcal{L}_r(\mathcal{G}) + \mathcal{L}_r(\mathcal{G})^T) \right) Q\beta}{\beta^T Q \beta} \\
&= \min_{\substack{\zeta \neq 0 \\ \zeta \perp \mathbf{1}}} \frac{\zeta^T \overline{\mathcal{L}}_r(\mathcal{G}) \zeta}{\zeta^T \zeta} = \mu_2(\overline{\mathcal{L}}_r(\mathcal{G})),
\end{aligned}$$

where β is an arbitrary vector with the appropriate dimension, Q is the matrix introduced in (14), and $Q^2 = Q$. In the last variational statement, we observe that ζ

should have a special structure, i.e., $\zeta = Q\beta$ (a zero at the row corresponding to the leader). An examination of the error dynamics suggests that such a structure always exists. As the leader does not update its value in the static leader case, the difference between the leader's state and the agreement value is always zero. Thus with respect to the disagreement dynamics (16),

$$\begin{aligned}\dot{V}(\zeta) &= -\zeta^T \,\overline{\mathcal{L}}_r(\mathcal{G})\,\zeta \leq -\mu_2(\mathcal{L}_r(\mathcal{G}))\zeta^T\zeta \\ &\leq -\lambda_2(\mathcal{L}(\mathcal{G}))\,\zeta^T\zeta. \qquad \square\end{aligned}$$

7. Controllability of multiple-leader networks. Some applications of multiagent systems may require multiple leaders. As our subsequent discussion shows, in this case, one needs an additional set of graph-theoretic tools to analyze the network controllability. In this venue, we first introduce equitable partitions and interlacing theory that play important roles in our analysis. We then present the main theorem of this section, providing a graph-theoretic characterization of controllability for multiple-leader networks.

7.1. Interlacing and equitable partitions. A *cell* $C \subset \mathcal{V}_\mathcal{G}$ is a subset of the node set. A *partition* of the graph is then a grouping of its node set into different cells.

DEFINITION 7.1. *An r-partition π of $\mathcal{V}_\mathcal{G}$, with cells $C_1, \ldots, C_r$, is said to be* equitable *if each node in C_j has the same number of neighbors in C_i for all i, j. We denote the cardinality of the partition π by $r = |\pi|$.*

Let b_{ij} be the number of neighbors in C_j of a node in C_i. The directed graph with the cells of an equitable r-partition π as its nodes, and with b_{ij} edges from the ith to the jth cells of π, is called the quotient of $\mathcal{G}$ over π and is denoted by $\mathcal{G}/\pi$. An obvious trivial partition is the n-partition, $\pi = \{\{1\}, \{2\}, \ldots, \{n\}\}$. If an equitable partition contains at least one cell with more than one node, we call it a nontrivial equitable partition (NEP), and the adjacency matrix of a quotient is given by

$$\mathcal{A}(\mathcal{G}/\pi)_{ij} = b_{ij}.$$

Equitable partitions of a graph can be obtained from its automorphisms. For example, in the Peterson graph shown in Figure 5(a), one equitable partition π_1 (Figure 5(b)) is given by the two orbit of the automorphism groups, namely the 5 inner vertices and the 5 outer vertices. The adjacency matrix of the quotient is then given by

$$\mathcal{A}(\mathcal{G}/\pi_1) = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

The equitable partition can also be introduced by the equal distance partition. Let $C_1 \subset \mathcal{V}_\mathcal{G}$ be a given cell, and let $C_i \subset \mathcal{V}_\mathcal{G}$ be the set of vertices at distance $i-1$ from C_1. C_1 is said to be *completely regular* if its distance partition is equitable. For instance, every node in the Peterson graph is completely regular and introduces the partition π_2 as shown in Figure 5(c). The adjacency matrix of this quotient is given by

$$\mathcal{A}(\mathcal{G}/\pi_2) = \begin{bmatrix} 0 & 3 & 0 \\ 1 & 0 & 2 \\ 0 & 1 & 2 \end{bmatrix}.$$

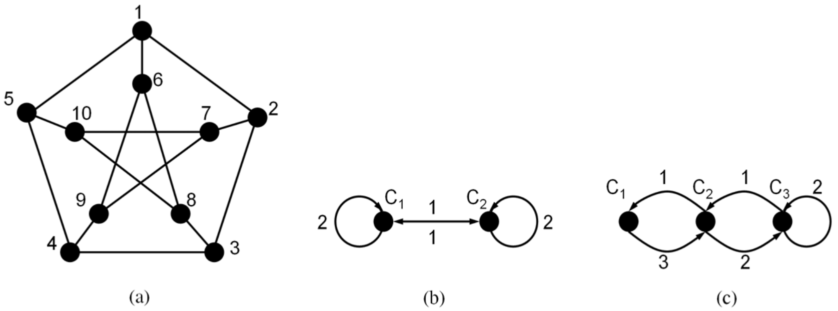


Fig. 5. (a) *Example of equitable partitions on the Peterson graph $\mathcal{G} = J(5,2,0)$ and the quotients;* (b) *the NEP introduced by the automorphism is $\pi_1 = \{C_1^1, C_2^1\}$, $C_1^1 = \{1,2,3,4,5\}$, $C_2^1 = \{6,7,8,9,10\}$; and* (c) *the NEP introduced by an equal-distance partition is $\pi_2 = \{C_1^2, C_2^2, C_3^2\}$, $C_1^2 = \{1\}$, $C_2^2 = \{2,5,6\}$, $C_3^2 = \{3,4,7,8,9,10\}$.*

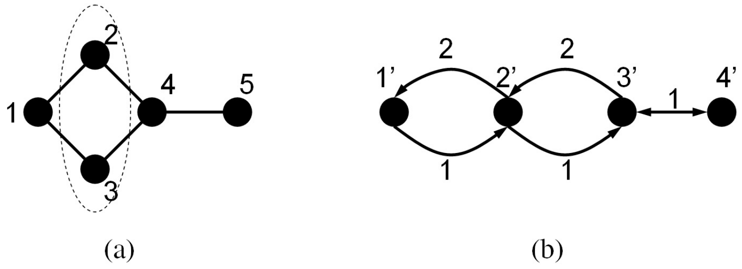


Fig. 6. (a) *The equitable partition and* (b) *the quotient of a graph.*

The adjacency matrix of the original graph and the quotient are closely related through the interlacing theorem. First, let us introduce the notion of the characteristic matrix of an equitable partition.

Definition 7.2. *A* characteristic vector $p_i \in \mathbb{R}^n$ *of a nontrivial cell C_i has 1's in components associated with C_i and 0's elsewhere.*[3] *A* characteristic matrix $P \in \mathbb{R}^{n \times r}$ *of a partition π of $\mathcal{V}_\mathcal{G}$ is a matrix with characteristic vectors of the cells as its columns. For example, the characteristic matrix of the equitable partition of the graph in Figure* 6(a) *is given by*

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{28}$$

Lemma 7.3 (see [14, Lemma 9.3.1]). *Let P be the characteristic matrix of an equitable partition π of the graph $\mathcal{G}$, and let $\hat{\mathcal{A}} = \mathcal{A}(\mathcal{G}/\pi)$. Then $\mathcal{A}P = P\hat{\mathcal{A}}$ and $\hat{\mathcal{A}} = P^+\mathcal{A}P$, where $P^+ = (P^TP)^{-1}P^T$ is the pseudo-inverse of P.*

[3] A nontrivial cell is a cell with more than one node.

As an example, the graph in Figure 6 has a nontrivial cell $(2, 3)$. The adjacency matrix of the original graph is

$$\mathcal{A} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

The adjacency matrix of the quotient, on the other hand, is

$$\hat{\mathcal{A}} = P^{+}\mathcal{A}P = \begin{bmatrix} 0 & 2 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 2 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

Lemma 7.4 (see [14, Lemma 9.3.2]). *Let $\mathcal{G}$ be a graph with adjacency matrix $\mathcal{A}$, and let π be a partition of $\mathcal{V}_{\mathcal{G}}$ with characteristic matrix P. Then π is equitable if and only if the column space of P is $\mathcal{A}$-invariant.*

Lemma 7.5 (see [23]). *Given a symmetric matrix $A \in \mathbb{R}^{n\times n}$, let S be a subspace of $\mathbb{R}^n$. Then $S^\perp$ is A-invariant if and only if S is A-invariant.*

The proof of this lemma is well known and can be found, for example, in [23].

Remark 7.6. Let $\mathcal{R}(\cdot)$ denote the range space. Suppose $|\mathcal{V}_{\mathcal{G}}| = n$, $|C_i| = n_i$, and $|\pi| = r$. Then we can find an orthogonal decomposition of $\mathbb{R}^n$ as

$$\mathbb{R}^n = \mathcal{R}(P) \oplus \mathcal{R}(Q). \tag{29}$$

In this case the matrix Q satisfies $\mathcal{R}(Q) = \mathcal{R}(P)^\perp$, and its columns, together with those of P, form a basis for $\mathbb{R}^n$. Note that by Lemma 7.5, $\mathcal{R}(Q)$ is also $\mathcal{A}$-invariant.

One way of obtaining the Q matrix is via the orthonormal basis of $\mathcal{R}(P)^\perp$. Let us denote the normalized matrix (each column of which is a norm one vector) by $\bar{Q}$. Next, define

$$\bar{P} = P(P^TP)^{-\frac{1}{2}} \tag{30}$$

as the normalized P matrix.[4] Since $\bar{P}$ and $\bar{Q}$ have the same column space as P and Q, respectively, they satisfy $\bar{P}^T\bar{Q} = \mathbf{0}$ and $\bar{Q}^T\bar{Q} = I_{n-r}$. In other words,

$$T = [\bar{P} \mid \bar{Q}] \tag{31}$$

is a matrix, constructed based on the equitable partition π, whose columns constitute an orthonormal basis for $\mathbb{R}^n$.

Theorem 7.7 (see [14, Theorem 9.3.3]). *If π is an equitable partition of a graph $\mathcal{G}$, then the characteristic polynomial of $\hat{\mathcal{A}} = \mathcal{A}(\mathcal{G}/\pi)$ divides the characteristic polynomial of $\mathcal{A}(\mathcal{G})$.*

Lemma 7.8 (see [14, Theorem 9.5.1]). *Let $\Phi \in \mathbb{R}^{n\times n}$ be a real symmetric matrix, and let $R \in \mathbb{R}^{n\times m}$ be such that $R^TR = I_m$. Set $\Theta = R^T\Phi R$ and let $\nu_1, \nu_2, \dots, \nu_m$ be an orthogonal set of eigenvectors for Θ such that $\Theta\nu_i = \lambda_i(\Theta)\nu_i$, where $\lambda_i(\Theta) \in \mathbb{R}$ is*

[4] Note that the invertibility of P^TP follows from the fact that the cells of the partition are nonempty. In fact, P^TP is a diagonal matrix with $(P^TP)_{ii} = |C_i|$.

an eigenvalue of Θ. *Then*

1. *the eigenvalues of* Θ *interlace the eigenvalues of* Φ.
2. *if* $\lambda_i(\Theta) = \lambda_i(\Phi)$, *then there is an eigenvector* v *of* Θ *with eigenvalue* $\lambda_i(\Theta)$ *such that* $R\nu$ *is an eigenvector of* Φ *with eigenvalue* $\lambda_i(\Phi)$.
3. *if* $\lambda_i(\Theta) = \lambda_i(\Phi)$ *for* $i = 1, \dots, l$, *then* $R\nu_i$ *is an eigenvector for* Φ *with eigenvalue* $\lambda_i(\Phi)$ *for* $i = 1, \dots, l$.
4. *if the interlacing is tight, then* $\Phi R = R\Theta$.

Based on the controllability results introduced in section 5, together with some basic properties of the graph Laplacian, we first derive the following lemma.

LEMMA 7.9. *Given a connected graph, the system* (7) *is controllable if and only if* $\mathcal{L}$ *and* $\mathcal{L}_f$ *do not share any common eigenvalues.*

Proof. We can reformulate the lemma as stating that the system is uncontrollable if and only if there exists at least one common eigenvalue between $\mathcal{L}$ and $\mathcal{L}_f$.

Necessity. Suppose that the system is uncontrollable. Then by Proposition 5.1 there exists a vector $\nu_i \in \mathbb{R}^{n_f}$ such that $\mathcal{L}_f\nu_i = \lambda\nu_i$ for some $\lambda \in \mathbb{R}$, with $l_{fl}^T\nu_i = \mathbf{0}$. Now, since

$$\begin{bmatrix} \mathcal{L}_f & l_{fl} \\ l_{fl}^T & \mathcal{L}_l \end{bmatrix} \begin{bmatrix} \nu_i \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathcal{L}_f\nu_i \\ l_{fl}^T\nu_i \end{bmatrix} = \lambda \begin{bmatrix} \nu_i \\ \mathbf{0} \end{bmatrix},$$

λ is also an eigenvalue of $\mathcal{L}$, with eigenvector $[\nu_i^T, \mathbf{0}]^T$. The necessary condition thus follows.

Sufficiency. It suffices to show that if $\mathcal{L}$ and $\mathcal{L}_f$ share a common eigenvalue, then the system $(\mathcal{L}, l_{fl})$ is not completely controllable. Since $\mathcal{L}_f$ is a principal submatrix of $\mathcal{L}$, it can be given by

$$\mathcal{L}_f = P_f^T \mathcal{L} P_f,$$

where $P_f = [I_{n_f}, 0]^T$ is the $n \times n_f$ matrix defined in (12). Following the fourth statement of Lemma 7.8,[5] if $\mathcal{L}_f$ and $\mathcal{L}$ share a common eigenvalue, say λ, then the corresponding eigenvector satisfies

$$\nu = P_f\nu_f = \begin{bmatrix} \nu_f \\ \mathbf{0} \end{bmatrix},$$

where ν is λ's eigenvector of $\mathcal{L}$ and ν_f is that of $\mathcal{L}_f$. Moreover, we know that

$$\mathcal{L}\nu = \begin{bmatrix} \mathcal{L}_f & l_{fl} \\ l_{fl}^T & \mathcal{L}_l \end{bmatrix} \begin{bmatrix} \nu_f \\ \mathbf{0} \end{bmatrix} = \lambda \begin{bmatrix} \nu_f \\ \mathbf{0} \end{bmatrix},$$

which gives us $l_{fl}^T\nu_f = \mathbf{0}$; thus the system is uncontrollable. □

Remark 7.10. Lemma 7.9 is an extension of Corollary 5.2, Propositions 5.3, and Proposition 5.4 to multileader settings.

7.2. Controllability analysis based on equitable partitions. In this section, we will utilize a graph-theoretic approach to characterize the necessary condition for a multiple-leader networked system to be controllable. The way we approach this necessary condition is through Lemma 7.9. In what follows we will show first that matrices $\mathcal{L}$ and $\mathcal{L}_f$ are both similar to some block diagonal matrices. Furthermore,

[5] Here the matrix P_f plays the same role as the matrix R in the fourth statement of Lemma 7.8.

we show that under certain assumptions, the diagonal block matrices obtained from the diagonalization of $\mathcal{L}$ and $\mathcal{L}_f$ have common diagonal block(s).

LEMMA 7.11. *If a graph $\mathcal{G}$ has an NEP π with characteristic matrix P, then the corresponding adjacency matrix $\mathcal{A}(\mathcal{G})$ is similar to a block diagonal matrix*

$$\bar{\mathcal{A}} = \left[\begin{array}{cc} \mathcal{A}_P & \mathbf{0} \\ \mathbf{0} & \mathcal{A}_Q \end{array} \right],$$

where $\mathcal{A}_P$ is similar to the adjacency matrix $\hat{\mathcal{A}} = \mathcal{A}(\mathcal{G}/\pi)$ of the quotient graph.

Proof. Let the matrix $T = [\bar{P} \mid \bar{Q}]$ be the orthonormal matrix with respect to π, as defined in (31). Let

$$\bar{\mathcal{A}} = T^T \mathcal{A} T = \left[\begin{array}{cc} \bar{P}^T \mathcal{A} \bar{P} & \bar{P}^T \mathcal{A} \bar{Q} \\ \bar{Q}^T \mathcal{A} \bar{P} & \bar{Q}^T \mathcal{A} \bar{Q} \end{array} \right]. \tag{32}$$

Since $\bar{P}$ and $\bar{Q}$ have the same column spaces as P and Q, respectively, they inherit their $\mathcal{A}$-invariance property, i.e., there exist matrices B and C such that

$$\mathcal{A}\bar{P} = \bar{P}B \quad \text{and} \quad \mathcal{A}\bar{Q} = \bar{Q}C.$$

Moreover, since the column spaces of $\bar{P}$ and $\bar{Q}$ are orthogonal complements of each other, one has

$$\bar{P}^T \mathcal{A} \bar{Q} = \bar{P}^T \bar{Q} C = \mathbf{0}$$

and

$$\bar{Q}^T \mathcal{A} \bar{P} = \bar{Q}^T \bar{P} B = \mathbf{0}.$$

In addition, by letting $D_p^2 = P^T P$, we obtain

$$\bar{P}^T \mathcal{A} \bar{P} = D_P^{-1} P^T \mathcal{A} P D_P^{-1} = D_P (D_P^{-2} P^T \mathcal{A} P) D_P^{-1} = D_P \hat{\mathcal{A}} D_P^{-1}, \tag{33}$$

and therefore the first diagonal block is similar to $\hat{\mathcal{A}}$. □

LEMMA 7.12. *Let P be the characteristic matrix of an NEP in $\mathcal{G}$. Then $\mathcal{R}(P)$ is K-invariant, where K is any diagonal block matrix of the form*

$$K = \mathbf{Diag}([\underbrace{k_1, \ldots, k_1}_{n_1}, \underbrace{k_2, \ldots, k_2}_{n_2}, \ldots, \underbrace{k_r, \ldots, k_r}_{n_r}]^T) = \mathbf{Diag}([k_i \mathbf{1}_{n_i}]_{i=1}^r),$$

$k_i \in \mathbb{R}$, $n_i = |C_i|$ is the cardinality of the cell, and $r = |\pi|$ is the cardinality of the partition. Consequently,

$$\bar{Q}^T K \bar{P} = \mathbf{0},$$

where $\bar{P} = P(P^T P)^{-\frac{1}{2}}$ and $\bar{Q}$ is chosen in such a way that $T = [\bar{P} \mid \bar{Q}]$ is an orthonormal matrix.

Proof. We note that

$$P = \left[\begin{array}{c} P_1 \\ P_2 \\ \vdots \\ P_r \end{array} \right] = \left[\begin{array}{cccc} p_1 & p_2 & \ldots & p_r \end{array} \right],$$

where $P_i \in \mathbb{R}^{n_i \times r}$ is a row block which has 1's in column i and 0's elsewhere. On the other hand, p_i is a characteristic vector representing C_i, which has 1's in the positions associated with C_i and zeros otherwise. Recall the example given in (28) with

$$P = \left[\begin{array}{c|c|cc} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right]; \tag{34}$$

we can then find

$$P_2 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix},$$

while $p_2 = [0\ 1\ 1\ 0\ 0]^T$. A little algebra reveals that

$$KP = \begin{bmatrix} k_1 P_1 \\ k_2 P_2 \\ \vdots \\ k_r P_r \end{bmatrix} = \begin{bmatrix} k_1 p_1 & k_2 p_2 & \dots & k_r p_r \end{bmatrix} = P\hat{K},$$

where $\hat{K} = \mathbf{Diag}([k_1, k_2, \dots, k_r]^T) = \mathbf{Diag}([k_i]_{i=1}^r)$; hence $\mathcal{R}(P)$ is K-invariant. Since $\mathcal{R}(\bar{Q}) = \mathcal{R}(P)^\perp$, by Lemma 7.5 it is K-invariant as well and

$$\bar{Q}^T K \bar{P} = \bar{Q}^T \bar{P} \hat{K} = \mathbf{0}. \qquad \square$$

By the definition of equitable partitions, the subgraph induced by a cell is regular and every node in the same cell has the same number of neighbors outside the cell. Therefore, the nodes belonging to the same cell have the same degree, and thus by Lemma 7.12, $\mathcal{R}(\bar{Q})$ and $\mathcal{R}(P)$ are $\mathcal{D}$-invariant, where $\mathcal{D}$ is the degree matrix given by

$$\mathcal{D} = \mathbf{Diag}([d_i \mathbf{1}_{n_i}]_{i=1}^r),$$

with $d_i \in \mathbb{R}$ denoting the degree of each node in the cell. Since the graph Laplacian satisfies $\mathcal{L}(\mathcal{G}) = \mathcal{D}(\mathcal{G}) - \mathcal{A}(\mathcal{G})$, Lemmas 7.11 and 7.12 imply that $\mathcal{R}(\bar{Q})$ and $\mathcal{R}(P)$ are $\mathcal{L}$-invariant. Thereby, we have following corollary.

COROLLARY 7.13. *Given the same condition as in Lemma* 7.11*, $\mathcal{L}$ is similar to a diagonal block matrix*

$$\bar{\mathcal{L}} = T^T \mathcal{L} T = \begin{bmatrix} \mathcal{L}_P & \mathbf{0} \\ \mathbf{0} & \mathcal{L}_Q \end{bmatrix}, \tag{35}$$

where $\mathcal{L}_P = \bar{P}^T \mathcal{L} \bar{P}$ and $\mathcal{L}_Q = \bar{Q}^T \mathcal{L} \bar{Q}$, and $T = [\bar{P} \mid \bar{Q}]$ defines an orthonormal basis for $\mathbb{R}^n$ with respect to π.

As (35) defines a similarity transformation, it follows that $\mathcal{L}_P$ and $\mathcal{L}_Q$ carry all the spectral information of $\mathcal{L}$, i.e., they share the same eigenvalues as $\mathcal{L}$.

As we have shown in section 2, in a leader-follower network, the graph Laplacian can be partitioned as

$$\mathcal{L} = \begin{bmatrix} \mathcal{L}_f & l_{fl} \\ l_{fl}^T & \mathcal{L}_l \end{bmatrix}.$$

Transformations similar to (35) can also be found for $\mathcal{L}_f$ in the presence of NEPs in the follower graph $\mathcal{G}_f$.

COROLLARY 7.14. *Let $\mathcal{G}_f$ be a follower graph, and let $\mathcal{L}_f$ be the diagonal submatrix of $\mathcal{L}$ related to $\mathcal{G}_f$. If there is an NEP π_f in $\mathcal{G}_f$ and a π in $\mathcal{G}$ such that all the nontrivial cells in π_f are also cells in π, then there exists an orthonormal matrix T_f such that*

$$\bar{\mathcal{L}}_f = T_f^T \mathcal{L}_f T_f = \begin{bmatrix} \mathcal{L}_{fP} & \mathbf{0} \\ \mathbf{0} & \mathcal{L}_{fQ} \end{bmatrix}. \tag{36}$$

Proof. Let $\bar{P}_f = P_f(P_f^T P_f)^{\frac{1}{2}}$, where P_f is the characteristic matrix for π_f. Moreover, let $\bar{Q}_f$ be defined on an orthonormal basis of $\mathcal{R}(P_f)^\perp$. In this way, we obtain an orthonormal basis for $\mathbb{R}^{n_f}$ with respect to π_f. Moreover, by (9), $\mathcal{L}_f(\mathcal{G}) = \mathcal{D}_f^l(\mathcal{G}) + \mathcal{L}(\mathcal{G}_f)$, where $\mathcal{L}(\mathcal{G}_f)$ denotes the Laplacian matrix of $\mathcal{G}_f$ while $\mathcal{D}_f^l$ is the diagonal follower-leader degree matrix defined in (8). Since all the nontrivial cells in π_f are also cells in π, $\mathcal{D}_f$ satisfies the condition in Lemma 7.12, i.e., nodes from an identical cell in π_f have the same degree. Hence by Lemma 7.11 and Lemma 7.12, $\mathcal{R}(\bar{P}_f)$ and $\mathcal{R}(\bar{Q}_f)$ are $\mathcal{L}_f$-invariant and consequently,

$$\bar{\mathcal{L}}_f = T_f^T \mathcal{L}_f T_f = \begin{bmatrix} \mathcal{L}_{fP} & \mathbf{0} \\ \mathbf{0} & \mathcal{L}_{fQ} \end{bmatrix}, \tag{37}$$

where $T_f = [\bar{P}_f \mid \bar{Q}_f]$, $\mathcal{L}_{fP} = \bar{P}_f^T \mathcal{L}_f \bar{P}_f$, and $\mathcal{L}_{fQ} = \bar{Q}_f^T \mathcal{L}_f \bar{Q}_f$. □

Again, the diagonal blocks of $\bar{\mathcal{L}}_f$ contain the entire spectral information of $\mathcal{L}_f$. We are now in the position to prove the main result of this section.

THEOREM 7.15. *Given a connected graph $\mathcal{G}$ and the induced follower graph $\mathcal{G}_f$, the system* (7) *is not controllable if there exist NEPs on $\mathcal{G}$ and $\mathcal{G}_f$, say π and π_f, such that all nontrivial cells of π are contained in π_f; i.e., for all $C_i \in \pi \backslash \pi_f$, one has $|C_i| = 1$.*

Proof. In Corollaries 7.13 and 7.14, we have shown that $\mathcal{L}$ and $\mathcal{L}_f$ are similar to some block diagonal matrices. Here we further expand on the relationship between such matrices.

Assume that $\pi \cap \pi_f = \{C_1, C_2, \ldots, C_{r_1}\}$. According to the underlying condition, one has $|C_i| \geq 2,\ i = 1, 2, \ldots, r_1$. Without loss of generality, we can index the nodes in such a way that the nontrivial cells comprise the first n_1 nodes, where[6]

$$n_1 = \sum_{i=1}^{r_1} |C_i| \leq n_f < n.$$

As all the nontrivial cells of π are in π_f, their characteristic matrices have similar structures,

$$P = \begin{bmatrix} P_1 & \mathbf{0} \\ \mathbf{0} & I_{n-n_1} \end{bmatrix}_{n \times r} \quad \text{and} \quad P_f = \begin{bmatrix} P_1 & \mathbf{0} \\ \mathbf{0} & I_{n_f-n_1} \end{bmatrix}_{n_f \times r_f},$$

where P_1 is an $n_1 \times r_1$ matrix containing the nontrivial part of the characteristic matrices. Since $\bar{P}$ and $\bar{P}_f$ are the normalizations of P and P_f, respectively, they

[6] We have introduced n_1 for notational convenience. It is easy to verify that $n_1 - r_1 = n - r = n_f - r_f$.

have the same block structures. Consequently $\bar{Q}$ and $\bar{Q}_f$, the matrices containing the orthonormal bases of $\mathcal{R}(P)$ and $\mathcal{R}(P_f)$, have the following structures:

$$\bar{Q} = \begin{bmatrix} Q_1 \\ \mathbf{0} \end{bmatrix}_{n\times(n_1-r_1)} \quad \text{and} \quad \bar{Q}_f = \begin{bmatrix} Q_1 \\ \mathbf{0} \end{bmatrix}_{n_f\times(n_1-r_1)},$$

where Q_1 is an $n_1 \times (n_1 - r_1)$ matrix that satisfies $Q_1^T P_1 = \mathbf{0}$. We observe that $\bar{Q}_f$ is different from $\bar{Q}$ only by $n - n_f$ rows of zeros. In other words, the special structures of $\bar{Q}$ and $\bar{Q}_f$ lead to the relationship

$$Q_f = R^T Q,$$

where $R = [I_{n_f}, 0]^T$. Now, recall the definition of $\mathcal{L}_Q$ and $\mathcal{L}_{Qf}$ from (35) and (36), leading us to

$$\mathcal{L}_Q = \bar{Q}^T \mathcal{L} \bar{Q} = \bar{Q}_f^T R^T \mathcal{L} R \bar{Q}_f = \bar{Q}_f^T \mathcal{L}_f \bar{Q}_f = \mathcal{L}_{fQ}. \tag{38}$$

Therefore $\mathcal{L}_f$ and $\mathcal{L}$ share the same eigenvalues associated with $\mathcal{L}_Q$; hence by Lemma 7.9, the system is not controllable. □

Theorem 7.15 provides a method to identify uncontrollable multi-agent systems in the presence of multiple leaders. In an uncontrollable multi-agent system, vertices in the same cell of an NEP, satisfying the condition in Theorem 7.15, are not distinguishable from the leaders' point of view. In other words, agents belonging to a shared cell among π and π_f, when identically initialized, remain undistinguished to the leaders throughout the system evolution. Moreover, the controllable subspace for this multi-agent system can be obtained by collapsing all the nodes in the same cell into a single "meta-agent." However, since the NEPs may not be unique, as we have seen in the case of the Peterson graph, more work is required before a complete understanding of the intricate interplay between controllability and NEPs is obtained.

Two immediate ramifications of the above theorem are as follows.

Corollary 7.16. *Given a connected graph $\mathcal{G}$ with the induced follower graph $\mathcal{G}_f$, a necessary condition for* (7) *to be controllable is that no NEPs π and π_f, on $\mathcal{G}$ and $\mathcal{G}_f$, respectively, share a nontrivial cell.*

Corollary 7.17. *If $\mathcal{G}$ is disconnected, a necessary condition for* (7) *to be controllable is that all of its connected components are controllable.*

8. Simulation and discussions.

In this section we will explore controllable and uncontrollable leader-follower networks that are amenable to analysis via methods proposed in this paper.

Example 1 (single leader with symmetric followers). In Figure 6, if we choose node 5 as the leader, the symmetric pair $(2, 3)$ in the follower graph renders the network uncontrollable, as stated in [34]. The dimension of the controllable subspace is three, while there are four nodes in the follower group. This result can also be interpreted via Theorem 7.15, since the corresponding automorphisms introduce equitable partitions.

Example 2 (single leader with equal distance partitions). We have shown in Figure 5 that the Peterson graph has two NEPs. One is introduced by the automorphism group and the other (π_2) is introduced by the equal-distance partition. Based on π_2, if we choose node 1 as the leader, the leader-follower network ends up with a controllable subspace of dimension two. Since there are four orbits in the automorphism group,[7] this dimension pertains to the two-cell equal-distance partitions.[8]

[7] They are $\{2, 5, 6\}, \{7, 10\}, \{8, 9\}$, and $\{3, 4\}$.

[8] They are $\{2, 5, 6\}$ and $\{3, 4, 7, 8, 9, 10\}$.

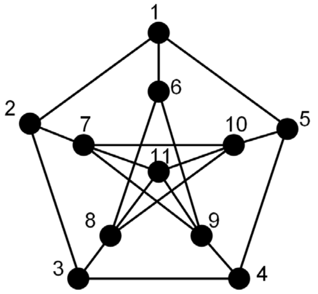


FIG. 7. *A* 2*-leader network based on the Peterson graph.*

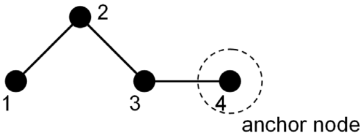


FIG. 8. *A path-like information exchange network.*

Example 3 (multiple leaders). This example is a modified leader graph based on the Peterson graph. In Figure 7, we add another node (11) connected to $\{3, 4, 7, 8, 9, 10\}$ as the second leader in addition to node 1. In this network, there is an equal-distance partition with four cells $\{1\}$, $\{2, 5, 6\}$, $\{3, 4, 7, 8, 9, 10\}$, and $\{11\}$. In this case, the dimension of the controllable subspace is still two, which is consistent with the second example above.

Example 4 (single-leader controllability). To demonstrate the controllability notion for the leader-follower system (7), consider a path-like information network, as shown in Figure 8. In this figure, the last node is chosen as the leader. By Proposition 5.17, this system is controllable. The system matrices in (7) assume the form

$$A = \begin{bmatrix} -1 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -2 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Using (26), one can find the controller that drives the leader-follower system from any initial state to an arbitrary final state. For this purpose, we chose to re-orient the planar triangle on the node set $\{1, 2, 3\}$. The maneuver time is set to be five seconds. Figure 9 shows the initial and the final positions of the nodes along with their respective trajectories.

Figure 10, on the other hand, depicts the leader node state trajectory as needed to perform the required maneuver. This trajectory corresponds to the speed of node 4 in the xy-plane. We note that as there are no restrictions on the leader's state trajectory, the actual implementation of this control law can become infeasible, especially when the maneuver time is arbitrarily short. This observation is apparent in the previous example, in this scenario, the speed of node 4 changes rather rapidly from 20 [m/s] to -50 [m/s]. To further explore the relationship between the location of the leader node and the convergence time to the agreement subspace, an extensive set of simulations was also carried out. In these simulations, at each step, a random connected graph

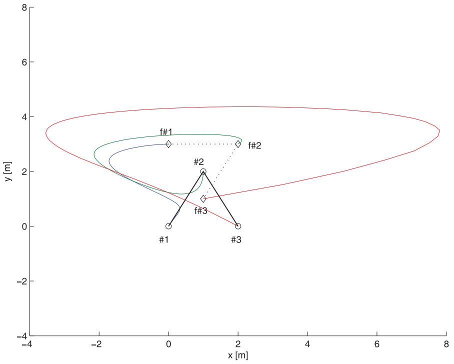


FIG. 9. *Initial and final positions of dynamic units and their respective state trajectories; $f\#i$ denotes the final position for agent i, $i = 1, 2, 3$.*

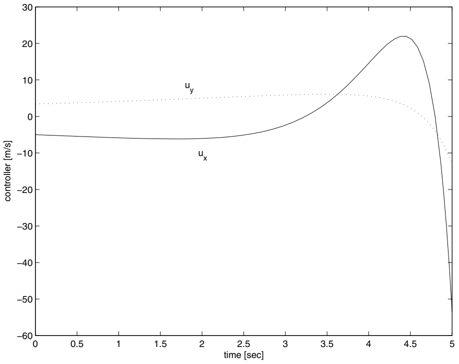


FIG. 10. *The leader node's velocity acts as a controller for the networked system.*

with 12 nodes and an edge probability of 0.3 was constructed. We then monitored the dynamics of the agreement protocol for the case when the center point of the graph was chosen to be the leader, as well as for the cases when the an arbitrary noncentral node is chosen.[9] These simulations were performed with 10 sets of randomly chosen initial conditions; the overall convergence time for each system was chosen to be the average of the total convergence times for all initial conditions. Figure 11 shows the result for 50 such iterations. We note that the convergence time is improved for the cases where the center of the graph is chosen as the leader.

9. Conclusions. In this paper, we considered the controlled agreement dynamics over a network. We first derived a set of transformations that can be employed to derive the system matrices for scenarios where one or more of the nodes (leader nodes) update their state values based on an external command. The other nodes in the graph (floating vertices) are assumed to update their states according to their relative states with their neighbors. In such a setting, we studied the controllability of the resulting dynamic system. It was shown that there is an intricate relationship between the uncontrollability of the corresponding multi-agent system and various graph-theoretic properties of the network. In particular, we pointed out the

[9] The center of the graph is a node with the following property: Its maximum distance to other nodes in the graph is minimum. We note that the center does not have to be unique.

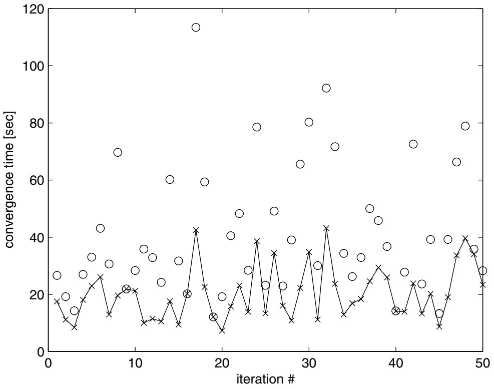


FIG. 11. *Convergence time comparison* (x: *center node is the leader.* o: *an arbitrary noncentral point is the leader*).

importance of the network automorphism group and its nontrivial equitable partitions in the controllability properties of the interconnected system. Some of the ramifications of this correspondence were then explored. The results of the present work point to a promising research direction at the intersection of graph theory and control theory that aims to study system-theoretic attributes from a purely graph-theoretic outlook.

Acknowledgments. The authors would like to thank the referees for their helpful comments and suggestions.

REFERENCES

[1] T. BALCH AND R. C. ARKIN, *Behavior-based formation control for multi-robot teams*, IEEE Trans. Robotics Automat., 14 (1998), pp. 926–939.

[2] B. BAMIEH, F. PAGANINI, AND M. DAHLEH, *Distributed control of spatially-invariant systems*, IEEE Trans. Automat. Control, 47 (2002), pp. 1091–1107.

[3] R. W. BEARD, J. R. LAWTON, AND F. Y. HADAEGH, *A coordination architecture for spacecraft formation control*, IEEE Trans. Control Systems Technology, 9 (2001), pp. 777–790.

[4] D. P. BERTSEKAS AND J. N. TSITSIKLIS, *Parallel and Distributed Computation*, Prentice–Hall, Englewood Cliffs, NJ, 1989.

[5] N. BIGGS, *Algebraic Graph Theory*, Cambridge University Press, Cambridge, UK, 1993.

[6] B.-D. CHEN AND S. LALL, *Dissipation inequalities for distributed systems on graphs*, in Proceedings of the 42nd IEEE Conference on Decision and Control, IEEE Press, Piscataway, NJ, 2003 pp. 3084–3090.

[7] J. CORTÉS AND F. BULLO, *Coordination and geometric optimization via distributed dynamical systems*, SIAM J. Control Optim., 44 (2005), pp. 1543–1574.

[8] J. CORTÉS, S. MARTÍNEZ, AND F. BULLO, *Robust rendezvous for mobile autonomous agents via proximity graphs in arbitrary dimensions*, IEEE Trans. Automat. Control, 8 (2006), pp. 1289–1298.

[9] R. D'ANDREA AND G. E. DULLERUD, *Distributed control design for spatially interconnected systems*, IEEE Trans. Automat. Control, 9 (2003), pp. 1478–1495.

[10] J. DESAI, J. OSTROWSKI, AND V. KUMAR, *Controlling formations of multiple mobile robots*, in Proceedings of the IEEE International Conference on Robotics and Automation, IEEE Press, Piscataway, NJ, 1998, pp. 2864–2869.

[11] R. DIESTEL, *Graph Theory*, Springer, New York, 2000.

[12] C. FALL, E. MARLAND, J. WAGNER, AND J. TYSON, EDS., *Computational Cell Biology*, Springer, New York, 2005.

[13] J. Fax and R. Murray, *Information flow and cooperative control of vehicle formations*, IEEE Trans. Automat. Control, 49 (2004), pp. 1465–1476.

[14] C. Godsil and G. Royle, *Algebraic Graph Theory*, Springer, New York, 2001.

[15] Y. Hatano and M. Mesbahi, *Agreement over random networks*, IEEE Trans. Automat. Control, 50 (2005), pp. 1867–1872.

[16] R. A. Horn and C. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.

[17] A. Jadbabaie, J. Lin, and A. S. Morse, *Coordination of groups of mobile autonomous agents using nearest neighbor rules*, IEEE Trans. Automat. Control, 6 (2003), pp. 988–1001.

[18] M. Ji, A. Muhammad, and M. Egerstedt, *Leader-based multi-agent coordination: Controllability and optimal control*, in Proceedings of the IEEE American Control Conference, IEEE Press, Piscataway, NJ, 2006, article 1656406.

[19] T. Kailath, *Linear Systems*, Prentice–Hall, Englewood Cliffs, NJ, 1980.

[20] Y. Kim, M. Mesbahi, and F. Y. Hadaegh, *Multiple-spacecraft reconfigurations through collision avoidance, bouncing, and stalemates*, J. Optim. Theory Appl., 2 (2004), pp. 323–343.

[21] J. Lauri and R. Scapellato, *Topics in Graph Automorphisms and Reconstruction*, Cambridge University Press, Cambridge, UK, 2003.

[22] Z. Lin, M. Broucke, and B. Francis, *Local control strategies for groups of mobile autonomous agents*, IEEE Trans. Automat. Control, 4 (2004), pp. 622–629.

[23] D. Luenberger, *Optimization by Vector Space Methods*, Wiley, New York, 1969.

[24] R. Merris, *Laplacian matrices of graphs: A survey*, Linear Algebra Appl., 197 (1994), pp. 143–176.

[25] M. Mesbahi, *State-dependent graphs*, in Proceedings of the 42nd IEEE Conference on Decision and Control, IEEE Press, Piscataway, NJ, 2003, pp. 3058–3063.

[26] M. Mesbahi, *State-dependent graphs and their controllability properties*, IEEE Trans. Automat. Control, 3 (2005), pp. 387–392.

[27] M. Mesbahi and F. Y. Hadaegh, *Formation flying control of multiple spacecraft via graphs, matrix inequalities, and switching*, J. Guidance Control Dynam., 2 (2001), pp. 369–377.

[28] L. Moreau, *Stability of multiagent systems with time-dependent communication links*, IEEE Trans. Automat. Control, 2 (2005), pp. 169–182.

[29] R. Olfati-Saber and R. M. Murray, *Agreement problems in networks with directed graphs and switching topology*, in Proceedings of the 42nd IEEE Conference on Decision and Control, IEEE Press, Piscataway, NJ, 2003, pp. 4126–4132.

[30] R. Olfati-Saber and R. M. Murray, *Consensus problems in networks of agents with switching topology and time-delays*, IEEE Trans. Automat. Control, 9 (2004), pp. 1520–1533.

[31] R. Olfati-Saber and J. S. Shamma, *Consensus filters for sensor networks and distributed sensor fusion*, in Proceedings of the 44th IEEE Conference on Decision and Control, IEEE Press, Piscataway, NJ, 2005, pp. 6698–6703.

[32] R. Olfati-Saber, *Flocking for multi-agent dynamic systems: Algorithms and theory*, IEEE Trans. Automat. Control, 3 (2006), pp. 401–420.

[33] K. Ogata, *Modern Control Engineering*, Prentice–Hall, Upper Saddle River, NJ, 2002.

[34] A. Rahmani and M. Mesbahi, *On the controlled agreement problem*, in Proceedings of the IEEE American Control Conference, IEEE Press, Piscataway, NJ, 2006, article 1656409.

[35] H. Tanner, A. Jadbabaie, and G. Pappas, *Stable flocking of mobile agents, part* II*: Dynamic topology*, in Proceedings of the 42nd IEEE Conference on Decision and Control, IEEE Press, Piscataway, NJ, 2003, pp. 2016–2021.

[36] H. G. Tanner, *On the controllability of nearest neighbor interconnections*, in Proceedings of the 43rd IEEE Conference on Decision and Control, IEEE Press, Piscataway, NJ, 2004, pp. 2467–2472.

[37] G. Walsh, H. Ye, and L. Bushnell, *Stability analysis of networked control systems*, in Proceedings of the American Control Conference, IEEE Press, Piscataway, NJ, 1999, pp. 2876–2880.

[38] P. K. C. Wang and F. Y. Hadaegh, *Coordination and control of multiple microspacecraft moving in formation*, J. Astronautical Sci., 44 (1996), pp. 315–355.

[39] L. Xiao and S. Boyd, *Fast linear iterations for distributed averaging*, Systems Control Lett., 53 (2004), pp. 65–78.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 1, pp. 187–205

© 2009 Society for Industrial and Applied Mathematics

MAINTAINING LIMITED-RANGE CONNECTIVITY AMONG SECOND-ORDER AGENTS[*]

KETAN SAVLA[†], GIUSEPPE NOTARSTEFANO[‡], AND FRANCESCO BULLO[§]

Abstract. In this paper we consider ad-hoc networks of robotic agents with double integrator dynamics. For such networks, the connectivity maintenance problems are as follows: (i) Do there exist control inputs for each agent to maintain network connectivity, and (ii) given desired controls for each agent, can we compute the closest connectivity-maintaining controls in a distributed fashion? The proposed solution is based on three contributions. First, we define and characterize admissible sets for double integrators to remain inside disks. Second, we establish an existence theorem for the connectivity maintenance problem by introducing a novel state-dependent graph, called the *double-integrator disk graph*. Specifically, we show that one can always maintain connectivity by maintaining a spanning tree of this new graph, but one will not always maintain connectivity of a particular agent pair that happens to be connected at one instant of time. Finally, we design a distributed "flow-control" algorithm for distributed computation of connectivity-maintaining controls.

Key words. multi-agent systems, connectivity maintenance, admissible sets, proximity graphs, distributed computation, solvability of linear inequalities

AMS subject classifications. 37N35, 68W15, 93C85, 93A14, 39B72

DOI. 10.1137/060674971

1. Introduction.

This work is a contribution to the emerging discipline of motion coordination for ad-hoc networks of mobile autonomous agents. This loose terminology refers to groups of robotic agents with limited mobility and communication capabilities. It is envisioned that such networks will perform a variety of useful tasks including surveillance, exploration, and environmental monitoring. The interest in this topic arises from the potential advantages of employing arrays of agents rather than single agents in certain applications. For example, from a control viewpoint, a group of agents inherently provides robustness to failures of single agents or of communication links.

The motion coordination problem for groups of autonomous agents is a control problem in the presence of communication constraints. Typically, each agent makes decisions based only on partial information about the state of the entire network that is obtained via communication with its immediate neighbors. One important difficulty is that the topology of the communication network depends on the agents' locations and, therefore, changes with the evolution of the network. In order to ensure a desired emergent behavior for a group of agents, it is necessary that the group does not disintegrate into subgroups that are unable to communicate with each other. In other words, some restrictions must be applied on the movement of the agents to

[*] Received by the editors November 15, 2006; accepted for publication (in revised form) August 4, 2008; published electronically February 11, 2009. This material is based upon work supported in part by the ARO MURI Award W911NF-05-1-0219, by the NSF Award CMS-0442041, and by the RECSYS Project IST-2001-37170. A preliminary version of this paper appears in *Proceedings of the* 2006 *American Control Conference* (Minneapolis, MN) [1].
http://www.siam.org/journals/sicon/48-1/67497.html

[†] Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139 (ksavla@mit.edu, http://web.mit.edu/ksavla/www).

[‡] Dipartimento di Ingegneria dell'Innovazione, Università del Salento, Via per Monteroni 73100 Lecce, Italy (giuseppe.notarstefano@unile.it, http://www.dei.unipd.it/∼notarste/).

[§] Center for Control, Dynamical Systems and Computation, University of California, Santa Barbara, CA 93106 (bullo@engineering.ucsb.edu, http://motion.mee.ucsb.edu).

ensure *connectivity* among the members of the group. In terms of design, it is required to constrain the control input such that the resulting topology maintains connectivity throughout its course of evolution. In [2], a connectivity constraint was developed for a group of agents modeled as first-order discrete time dynamic systems. In [2] and the related references [3, 4], this constraint is used to solve rendezvous problems. Connectivity constraints for line-of-sight communication are proposed in [5]. Another approach to connectivity maintenance for first-order systems is proposed in [6]. In [7], a centralized procedure for finding the set of control inputs that maintain k-hop connectivity for a network of agents is given. However, there is no guarantee that the resulting set of feasible control inputs is nonempty. In this paper we fully characterize the set of admissible control inputs for a group of agents modeled as second-order discrete time dynamic systems, which ensures connectivity of the group in the same spirit as described earlier.

The contributions of the paper are threefold. First, we consider a control system consisting of a double integrator with bounded control inputs. For such a system, we define and characterize the admissible set that allows the double integrator to remain inside disks. Second, we define a novel state-dependent graph—the *double-integrator disk graph*—and give an existence theorem for the connectivity maintenance problem for networks of second-order agents with respect to an appropriate version of this new graph. Specifically, we show that one can always maintain connectivity by maintaining a spanning tree through a subset of all edges without necessarily maintaining connectivity of any particular agent pair that happens to be connected at an instant of time. Remarkably, this conclusion is different from the one for the single integrator kinematic agent model. Finally, we formulate and solve an optimization problem for the distributed computation of connectivity-maintaining controls. Specifically, given a set of desired control inputs for all the agents, we aim to compute the set of connectivity-mantaining inputs that are "closest" to the desired ones. We set up this design problem as a standard quadratic programming problem and provide a distributed "flow-control" algorithm to solve it. As an example application, we solve this optimization problem for a particular set of desired controls to achieve a behavior reminiscent of the well-studied "flocking" behavior (see, e.g., [8]) among second-order agents with bounded controls while maintaining connectivity, something that has not been reported in the literature so far.

The paper is organized as follows. In section 2, we define and characterize the admissible sets for a double integrator to remain inside a disk, and based on this we define a new graph—the *double-integrator disk graph*. In section 3, we provide an existence theorem for the set of control inputs for the whole network of second-order agents that maintains connectivity with respect to an appropriately scaled version of this new graph. We also characterize and give an inner polytopic representation of the constraint set for these connectivity-maintaining control inputs. In section 4, we propose an optimization problem to compute connectivity-maintaining controls.

We also provide some illustrative simulations which suggest an alternative way of achieving a weak form of flocking of the agents. Finally, in section 5 we conclude with a few remarks about future work.

2. Preliminary developments. We begin with some notations. We let $\mathbb{N}$, $\mathbb{N}_0$, and $\mathbb{R}_+$ denote the natural numbers, the nonnegative integer numbers, and the positive real numbers, respectively. For $d \in \mathbb{N}$, we let 0_d and 1_d denote the vectors in $\mathbb{R}^d$ whose entries are all 0 and 1, respectively. We let $\|p\|$ denote the Euclidean norm of $p \in \mathbb{R}^d$. For $r \in \mathbb{R}_+$ and $p \in \mathbb{R}^d$, we let $B(p, r)$ denote the closed ball centered at

p with radius r, i.e., $B(p,r) = \{q \in \mathbb{R}^d \mid \|p-q\| \leq r\}$. For $x, y \in \mathbb{R}^d$, we let $x \preceq y$ denote component-wise inequality, i.e., $x_k \leq y_k$ for $k \in \{1,\dots,d\}$. We let $f : A \rightrightarrows B$ denote a set-valued map; in other words, for each $a \in A$, $f(a)$ is a subset of B. We identify $\mathbb{R}^d \times \mathbb{R}^d$ with $\mathbb{R}^{2d}$.

2.1. Maintaining a double integrator inside a disk. For $t \in \mathbb{N}_0$, consider the discrete time control system in $\mathbb{R}^{2d}$,

$$\begin{aligned} p[t+1] &= p[t] + v[t], \\ v[t+1] &= v[t] + u[t], \end{aligned} \tag{2.1}$$

where the norm of the control is upper-bounded by $r_{\mathrm{ctr}} \in \mathbb{R}_+$, i.e., $u[t] \in B(0_d, r_{\mathrm{ctr}})$ for $t \in \mathbb{N}_0$. We refer to this control system as the *discrete time double integrator* in $\mathbb{R}^d$ or, more loosely, as a second-order system. Given $(p,v) \in \mathbb{R}^{2d}$ and $\{u_\tau\}_{\tau \in \mathbb{N}_0} \subseteq B(0_d, r_{\mathrm{ctr}})$, let $\phi(t,(p,v),\{u_\tau\})$ denote the solution of (2.1) at time $t \in \mathbb{N}_0$ from initial condition (p,v) with inputs $u_1,\dots,u_{t-1}$.

In what follows we consider the following problem: Assume that the initial position of (2.1) is inside a disk centered at 0_d, then find inputs that keep it inside that disk. This task is impossible for general values of the initial velocity. In what follows we identify assumptions on the initial velocity that render this task possible.

For $r_{\mathrm{pos}} \in \mathbb{R}_+$, we define the *admissible set at time zero* by

$$\mathcal{A}_0^d(r_{\mathrm{pos}}) = B(0_d, r_{\mathrm{pos}}) \times \mathbb{R}^d.$$

For $r_{\mathrm{pos}}, r_{\mathrm{ctr}} \in \mathbb{R}_+$, we define the *admissible set for m time steps* by

$$\begin{aligned} \mathcal{A}_m^d(r_{\mathrm{pos}}, r_{\mathrm{ctr}}) = \big\{(p,v) \in \mathbb{R}^{2d} \mid \exists \{u_\tau\}_{\tau \in [0,m-1]} \subseteq B(0_d, r_{\mathrm{ctr}}) \\ \text{s.t. } \phi(t,(p,v),\{u_\tau\}) \in \mathcal{A}_0^d(r_{\mathrm{pos}}) \ \forall t \in [0,m]\big\}, \end{aligned}$$

and we define the *admissible set* by

$$\begin{aligned} \mathcal{A}^d(r_{\mathrm{pos}}, r_{\mathrm{ctr}}) = \big\{(p,v) \in \mathbb{R}^{2d} \mid \exists \{u_\tau\}_{\tau \in \mathbb{N}_0} \subseteq B(0_d, r_{\mathrm{ctr}}) \\ \text{s.t. } \phi(t,(p,v),\{u_\tau\}) \in \mathcal{A}_0^d(r_{\mathrm{pos}}) \ \forall t \in \mathbb{N}_0\big\}. \end{aligned}$$

With slight abuse of notation we shall sometimes suppress the arguments in the definitions of admissible sets. The following theorem establishes some important properties of the admissible sets.

THEOREM 2.1 (properties of the admissible sets). *For all $d \in \mathbb{N}$ and $r_{\mathrm{pos}}, r_{\mathrm{ctr}} \in \mathbb{R}_+$, the following statements hold:*

(i) *For all $m \in \mathbb{N}$, $\mathcal{A}_m^d(r_{\mathrm{pos}}, r_{\mathrm{ctr}}) \subseteq \mathcal{A}_{m-1}^d(r_{\mathrm{pos}}, r_{\mathrm{ctr}})$ and*

$$\mathcal{A}^d(r_{\mathrm{pos}}, r_{\mathrm{ctr}}) = \lim_{m \to +\infty} \mathcal{A}_m^d(r_{\mathrm{pos}}, r_{\mathrm{ctr}}) = \lim_{m \to +\infty} \cap_{k=1}^m \mathcal{A}_k^d(r_{\mathrm{pos}}, r_{\mathrm{ctr}})\,;$$

(ii) *$\mathcal{A}^d(r_{\mathrm{pos}}, r_{\mathrm{ctr}})$ is a convex, compact set and is the largest controlled-invariant [1] subset of $\mathcal{A}_0^d(r_{\mathrm{pos}})$;*

(iii) *$\mathcal{A}^d(r_{\mathrm{pos}}, r_{\mathrm{ctr}})$ is invariant under orthogonal transformations in the sense that, if $(p,v) \in \mathcal{A}^d(r_{\mathrm{pos}}, r_{\mathrm{ctr}})$, then also $(Rp, Rv) \in \mathcal{A}^d(r_{\mathrm{pos}}, r_{\mathrm{ctr}})$ for all orthogonal [2] matrices R in $\mathbb{R}^{d \times d}$;*

[1] A set is controlled invariant for a control system if there exists a feedback law such that the set is positively invariant for the closed-loop system.

[2] A matrix $R \in \mathbb{R}^{d \times d}$ is orthogonal if $RR^T = R^T R = I_d$.

(iv) *if* $0 < r_1 < r_2$, *then* $\mathcal{A}^d(r_{\text{pos}}, r_1) \subset \mathcal{A}^d(r_{\text{pos}}, r_2)$ *and* $\mathcal{A}^d(r_1, r_{\text{ctr}}) \subset \mathcal{A}^d(r_2, r_{\text{ctr}})$.

Proof. The two facts in statement (i) are direct consequences of the definitions of $\mathcal{A}^d_m$ and $\mathcal{A}^d$. Regarding statement (ii), each $\mathcal{A}^d_m$, $m \in \mathbb{N}$, is closed, the intersection of closed sets is closed, and, therefore, $\mathcal{A}^d = \lim_{m\to+\infty} \cap_{k=1}^m \mathcal{A}^d_k$ is closed. To show that $\mathcal{A}^d$ is bounded it suffices to show that $\mathcal{A}^d_1$ is bounded. Note that $(p, v) \in \mathcal{A}^d_1$ implies that there exists $u \in B(0_d, r_{\text{ctr}})$ such that $(p, v) \in \mathcal{A}^d_0$ and $(p+v, v+u) \in \mathcal{A}^d_0$. This, in turn, implies that $p \in B(0_d, r_{\text{pos}})$ and $p + v \in B(0_d, r_{\text{pos}})$. Therefore, $\mathcal{A}^d_1$ is bounded. Next, we prove that $\mathcal{A}^d_m$ is convex. Given (p_1, v_1) and (p_2, v_2) in $\mathcal{A}^d_m$, let u_1 and u_2 be controls that ensure that $\phi(t, (p_i, v_i), \{u_i\}) \in \mathcal{A}^d_0$, $t \in [0, m]$, $i \in \{1, 2\}$. For $\lambda \in [0, 1]$, consider the initial condition $(p_\lambda, v_\lambda) = (\lambda p_1 + (1-\lambda)p_2, \lambda v_1 + (1-\lambda)v_2)$ and the input $u_\lambda = \lambda u_1 + (1-\lambda)u_2$, and note that, by linearity,

$$\phi(t, (p_\lambda, v_\lambda), u_\lambda) = \lambda\phi(t, (p_1, v_1), \{u_1\}) + (1-\lambda)\phi(t, (p_2, v_2), \{u_2\}), \quad t \in [0, m].$$

Since $\phi(t, (p_1, v_1), \{u_1\})$ and $\phi(t, (p_2, v_2), \{u_2\})$ belong to the convex set $\mathcal{A}^d_0$, so does their convex combination. Therefore, (p_λ, v_λ) belongs to $\mathcal{A}^d_m$, and $\mathcal{A}^d_m$ is convex. Finally, $\mathcal{A}^d$ is convex because the intersection of convex sets is convex.

Next, we show that $\mathcal{A}^d$ is controlled invariant. Given $(p, v) \in \mathcal{A}^d$ (with corresponding control sequence $\{u_\tau\}_{\tau\in\mathbb{N}_0}$), we need to show that there exists a control input $x \in B(0_d, r_{\text{ctr}})$ such that $\phi(1, (p, v), x) \in \mathcal{A}^d$. Such an input can be chosen as $x = u_0$. Indeed, the control sequence $\{u_{\tau+1}\}_{\tau\in\mathbb{N}_0}$ keeps the trajectory starting from $\phi(1, (p, v), x)$ inside $\mathcal{A}^d_0$ and, therefore, $\phi(1, (p, v), x) \in \mathcal{A}^d$. Additionally, it is easy to see that $\mathcal{A}^d \subset \mathcal{A}^d_0$. Finally, we need to prove that $\mathcal{A}^d$ is the largest controlled-invariant subset of $\mathcal{A}^d_0$. Assume that there exists $\mathcal{A}^{d*}$ with the same properties and that is larger than $\mathcal{A}^d$. This means that there exists $(p, v) \in \mathcal{A}^{d*} \setminus \mathcal{A}^d$. This is equivalent to saying that $\exists\ \tau^* \in \mathbb{N}_0$ such that for every choice of the input u, $\phi(\tau^*, (p, v), u) \notin \mathcal{A}^d_0$. But, since $\mathcal{A}^{d*} \subset \mathcal{A}^d_0$, this leads to a contradiction.

Regarding statement (iii), observe that, if $(p, v) \in \mathcal{A}^d_0$, then $(Rp, Rv) \in \mathcal{A}^d_0$ and, if $u \in B(0, r_{\text{ctr}})$, then $Ru \in B(0, r_{\text{ctr}})$. Therefore, using again the linearity of the maps ϕ, the proof follows. Regarding statement (iv), the two results follow from the definition of $\mathcal{A}^d(r_{\text{pos}}, r_{\text{ctr}})$ and the facts that for all $0 < r_1 < r_2$, $B(0, r_1) \subset B(0, r_2)$ and $\mathcal{A}^d_0(r_1) \subset \mathcal{A}^d_0(r_2)$. □

Next, we study the set-valued map that associates to each state in $\mathcal{A}^d(r_{\text{pos}}, r_{\text{ctr}})$ the set of control inputs that keep the state inside $\mathcal{A}^d(r_{\text{pos}}, r_{\text{ctr}})$ in one step. We define the *admissible control set* $\mathcal{U}^d(r_{\text{pos}}, r_{\text{ctr}}) : \mathcal{A}^d(r_{\text{pos}}, r_{\text{ctr}}) \rightrightarrows B(0_d, r_{\text{ctr}})$ by

$$\mathcal{U}^d(r_{\text{pos}}, r_{\text{ctr}}) \cdot (p, v) = \{u \in B(0_d, r_{\text{ctr}}) \mid (p + v, v + u) \in \mathcal{A}^d(r_{\text{pos}}, r_{\text{ctr}})\},$$

or, equivalently,

$$\mathcal{U}^d(r_{\text{pos}}, r_{\text{ctr}}) \cdot (p, v) = B(0_d, r_{\text{ctr}}) \cap \{w - v \mid (p + v, w) \in \mathcal{A}^d(r_{\text{pos}}, r_{\text{ctr}})\}. \tag{2.2}$$

Lemma 2.2 (properties of the admissible control set). *For all* $(p, v) \in \mathcal{A}^d(r_{\text{pos}}, r_{\text{ctr}})$, *the set* $\mathcal{U}^d(r_{\text{pos}}, r_{\text{ctr}}) \cdot (p, v)$ *is nonempty, convex, and compact. For generic* $(p, v) \in \mathcal{A}^d(r_{\text{pos}}, r_{\text{ctr}})$, *the set* $\mathcal{U}^d(r_{\text{pos}}, r_{\text{ctr}}) \cdot (p, v)$ *does not contain* 0_d.

Proof. The nonemptiness of the set $\mathcal{U}^d(r_{\text{pos}}, r_{\text{ctr}}) \cdot (p, v)$ derives directly from the definition of $\mathcal{A}^d(r_{\text{pos}}, r_{\text{ctr}})$. Clearly, from (2.2), $\mathcal{U}^d(r_{\text{pos}}, r_{\text{ctr}}) \cdot (p, v)$ is closed (it is the intersection of two closed sets). It is also bounded ($\mathcal{U}^d(r_{\text{pos}}, r_{\text{ctr}}) \cdot (p, v) \subset B(0_d, r_{\text{ctr}})$); hence it is compact. To prove that it is convex, we need to show the following: Given $(p, v) \in \mathcal{A}^d(r_{\text{pos}}, r_{\text{ctr}})$, if there exist u_1 and u_2 in $\mathcal{U}^d(r_{\text{pos}}, r_{\text{ctr}}) \cdot (p, v)$ such that $\phi(1, (p, v), u_1)$ and $\phi(1, (p, v), u_2)$ belong to $\mathcal{A}^d(r_{\text{pos}}, r_{\text{ctr}})$, then $u_{12} = \lambda u_1 + (1-\lambda)u_2$,

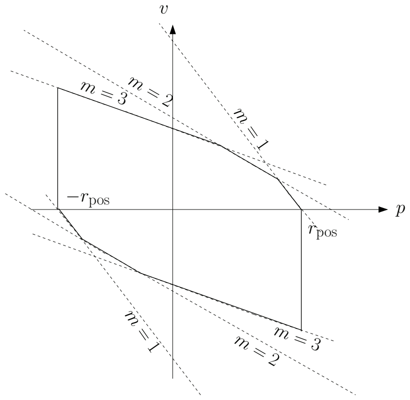


FIG. 2.1. *The admissible set* $\mathcal{A}^1$ *for generic values of* r_{pos} *and* r_{ctr}.

$\lambda \in [0,1]$, belongs to $\mathcal{U}^d(r_{\text{pos}}, r_{\text{ctr}}) \cdot (p,v)$, that is, $\phi(1,(p,v),u_{12}) \in \mathcal{A}^d(r_{\text{pos}}, r_{\text{ctr}})$. But this fact follows directly from the linearity of ϕ and the convexity of $\mathcal{A}^d(r_{\text{pos}}, r_{\text{ctr}})$. This proves that $\mathcal{U}^d(r_{\text{pos}}, r_{\text{ctr}}) \cdot (p,v)$ is convex. The fact that it does not necessarily contain the origin can be proved by contradiction as follows. Consider a $(p,v) \in \mathcal{A}^d(r_{\text{pos}}, r_{\text{ctr}})$ such that $v \neq 0_d$ and $\mathcal{U}^d(r_{\text{pos}}, r_{\text{ctr}}) \cdot (p,v)$ contains 0_d. This means that $(p+v,v)$ also belongs to $\mathcal{A}^d(r_{\text{pos}}, r_{\text{ctr}})$. Now, either $\mathcal{U}^d(r_{\text{pos}}, r_{\text{ctr}}) \cdot (p+v,v)$ does not contain 0_d, in which case we have proved the statement, or $\mathcal{A}^d(r_{\text{pos}}, r_{\text{ctr}})$ also contains $(p+2v,v)$. Continuing along these lines, if it were true that $\mathcal{U}^d(r_{\text{pos}}, r_{\text{ctr}}) \cdot (p,v)$ contains the origin for all $(p,v) \in \mathcal{A}^d(r_{\text{pos}}, r_{\text{ctr}})$, then one could show that $(p+tv,v)$ belongs to $\mathcal{A}^d(r_{\text{pos}}, r_{\text{ctr}})$ for all $t \in \mathbb{N}$. However, $\mathcal{A}^d(r_{\text{pos}}, r_{\text{ctr}})$ is bounded by Theorem 2.1. Hence, one can always find a $t^* \in \mathbb{N}$ such that $(p+t^*v,v) \in \mathcal{A}^d(r_{\text{pos}}, r_{\text{ctr}})$ but $(p+(t^*+1)v,v) \notin \mathcal{A}^d(r_{\text{pos}}, r_{\text{ctr}})$, thereby proving that $\mathcal{U}^d(r_{\text{pos}}, r_{\text{ctr}}) \cdot (p+t^*v,v)$ does not contain 0_d. □

2.2. Computing admissible sets. We characterize $\mathcal{A}^d$ for $d=1$ in the following result and we illustrate the outcome in Figure 2.1.

LEMMA 2.3 (admissible set in one dimension). *For* $r_{\text{pos}}, r_{\text{ctr}} \in \mathbb{R}_+$, *the following hold:*

(i) $\mathcal{A}^1(r_{\text{pos}}, r_{\text{ctr}})$ *is the polytope containing the points* $(p,v) \in \mathbb{R}^2$ *satisfying*

$$-\frac{r_{\text{pos}}}{m} - \frac{m-1}{2} r_{\text{ctr}} \leq v + \frac{p}{m} \leq \frac{r_{\text{pos}}}{m} + \frac{m-1}{2} r_{\text{ctr}} \tag{2.3}$$

for all $m \in \mathbb{N}$, *and* $p \in [-r_{\text{pos}}, r_{\text{pos}}]$;

(ii) *if* $\widehat{m}(r_{\text{pos}}, r_{\text{ctr}}) \in \mathbb{N}$ *is defined by*

$$\widehat{m}(r_{\text{pos}}, r_{\text{ctr}}) = \left\lceil -\frac{1}{2} + \sqrt{\frac{1}{4} + \frac{4r_{\text{pos}}}{r_{\text{ctr}}}} \right\rceil, \tag{2.4}$$

then $\mathcal{A}^1 = \mathcal{A}^1_m = \mathcal{A}^1_{\widehat{m}(r_{\text{pos}}, r_{\text{ctr}})}$, *for* $m \geq \widehat{m}(r_{\text{pos}}, r_{\text{ctr}})$.

Proof. Regarding statement (i), it suffices to show that, for $m \in \mathbb{N}$, $\mathcal{A}^1_m(r_{\text{pos}}, r_{\text{ctr}})$ is the set of points in $\mathcal{A}^1_{m-1}(r_{\text{pos}}, r_{\text{ctr}})$ that satisfy (2.3). If we show that this property holds for all m, then we can use statement (i) of Theorem 2.1 to complete the proof. Consider the set of equations (2.1) for m consecutive time indices after t. The solution of the linear system can be written in terms of the state at instant t as

$$\begin{bmatrix} p[t+m] \\ v[t+m] \end{bmatrix} = \begin{bmatrix} 1 & m \\ 0 & 1 \end{bmatrix} \begin{bmatrix} p[t] \\ v[t] \end{bmatrix} + \sum_{\tau=0}^{m-1} \begin{bmatrix} 1 & (m-1-\tau) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} u[t+\tau]. \tag{2.5}$$

It is clear that the points on the boundary of $\mathcal{A}^1_m$ have the property that the maximum control effort is needed to enforce the constraint. In other words we look for the points $(p[t], v[t]) \in \mathcal{A}^1_0$ with $v[t] \geq 0$ (the case $v[t] \leq 0$ can be solved in a similar way) such that the points $p[t+m] \leq r_{\text{cmm}}$ are reached by using the maximum control effort $u[t+\tau] = -r_{\text{ctr}}$, $\tau \in \{0, \dots, m-1\}$.

Substituting the expression of the control in (2.5), we obtain

$$p[t+m] = p[t] + mv[t] - r_{\text{ctr}} \sum_{\tau=0}^{m-1} (m-1-\tau), \quad v[t+m] = v[t] - mr_{\text{ctr}},$$

and using the equality $\sum_{\tau=0}^{m-1}(m-1-\tau) = \frac{m(m-1)}{2}$, we have

$$p[t+m] = p[t] + mv[t] - r_{\text{ctr}} \frac{m(m-1)}{2}, \quad v[t+m] = v[t] - mr_{\text{ctr}}. \tag{2.6}$$

In order to belong to $\mathcal{A}^1_m$, the point $(p[t], v[t])$ must satisfy the constraint $p[t+\tau] \leq r_{\text{cmm}}$, $\tau \in \{1, \dots, m\}$, or equivalently,

$$v[t] \leq -\frac{p[t]}{\tau} + \frac{r_{\text{cmm}}}{\tau} + r_{\text{ctr}} \frac{(\tau - 1)}{2}, \quad \tau \in \{1, \dots, m\}.$$

Using the same procedure for the points in the half plane $v[t] \leq 0$ (in this case the control is $u[t+\tau] = r_{\text{ctr}}$, $\tau \in \{0, \dots, m-1\}$), it turns out that $\mathcal{A}^1_m$ is equal to the set of all pairs $(p, v) \in \mathcal{A}^1_0$ satisfying

$$-\frac{p}{\tau} - \frac{r_{\text{cmm}}}{\tau} - \frac{\tau - 1}{2} r_{\text{ctr}} \leq v \leq -\frac{p}{\tau} + \frac{r_{\text{cmm}}}{\tau} + \frac{\tau - 1}{2} r_{\text{ctr}}, \quad \tau \in \{1, \dots, m\}.$$

By using statement (i) of Theorem 2.1, the proof is complete.

Regarding statement (ii), let us consider the case $v[t] \geq 0$ and evaluate the points on the boundary such that $(p[t+m], v[t+m]) = (r_{\text{cmm}}, 0)$, $m \in \mathbb{N}$. These points are obtained by substituting the above value of $(p[t+m], v[t+m])$ in (2.6). The points obtained are (p, v) such that

$$p = r_{\text{cmm}} - m \frac{(m+1)}{2} r_{\text{ctr}}, \qquad m \in \mathbb{N}_0.$$

It is easy to see that $\widehat{m}(r_{\text{pos}}, r_{\text{ctr}})$, as defined in (2.4), is the lowest m such that $p \leq -r_{\text{cmm}}$. □

Remarks 2.4.

(i) If $r_{\text{ctr}} \geq 2r_{\text{pos}}$, then $\mathcal{A}^1 = \mathcal{A}^1_1$; that is, for sufficiently large $r_{\text{ctr}}/r_{\text{pos}}$, the admissible set is equal to the admissible set in one time step.

(ii) The methodology for constructing $\mathcal{A}^1(r_{\text{pos}}, r_{\text{ctr}})$ is related to the procedure for constructing the so-called isochronic regions for discrete time optimal control of double integrators, as outlined in [9].

Next, we introduce some definitions useful for providing an inner approximation of $\mathcal{A}^d$ when $d \geq 2$. Given $p \in \mathbb{R}^d$ and $v \in \mathbb{R}^d \setminus \{0_d\}$, define $p_\| \in \mathbb{R}$ and $p_\perp \in \mathbb{R}^d$ by

$$p = p_\| \frac{v}{\|v\|} + p_\perp,$$

where $p_\perp \cdot v = 0$. For $r_{\text{pos}}, r_{\text{ctr}} \in \mathbb{R}_+$, define

$$\begin{aligned}(2.7) \quad \mathcal{A}_\|^d(r_{\text{pos}}, r_{\text{ctr}}) = \Big\{ &(p, v) \in B(0_d, r_{\text{pos}}) \times \mathbb{R}^d \mid v = 0_d \text{ or} \\ &(p_\|, \|v\|) \in \mathcal{A}^1 \left(\sqrt{r_{\text{pos}}^2 - \|p_\perp\|^2}, r_{\text{ctr}} \right) \Big\}.\end{aligned}$$

LEMMA 2.5. *For* $r_{\text{pos}}, r_{\text{ctr}} \in \mathbb{R}_+$, $\mathcal{A}_\|^d(r_{\text{pos}}, r_{\text{ctr}})$ *is a compact subset of* $\mathcal{A}^d(r_{\text{pos}}, r_{\text{ctr}})$.

Proof. We begin by showing that definition (2.7) is equivalent to

$$\begin{aligned}(2.8) \quad \mathcal{A}_\|^d(r_{\text{pos}}, r_{\text{ctr}}) = \Bigg\{ &(p, v) \in \mathcal{A}_0^d v = 0_d \text{ or } \exists \{u_{\|\tau}\}_{\tau \in \mathbb{N}_0} \subseteq [-r_{\text{ctr}}, r_{\text{ctr}}] \\ &\text{s.t. } \phi\left(t, (p, v), \{u_{\|\tau}\} \frac{v}{\|v\|}\right) \in \mathcal{A}_0^d(r_{\text{pos}}) \ \forall t \in \mathbb{N}_0 \Bigg\}.\end{aligned}$$

To establish this equivalence, we use the definition of the set $\mathcal{A}^1$. For $v \neq 0_d$, we rewrite the solution of the system as

$$\phi(t, (p, v), \{u_\tau\}) = \phi_\|(t, (p, v), \{u_\tau\}) \frac{v}{\|v\|} + \phi_\perp(t, (p, v), \{u_\tau\}),$$

where $\phi_\perp(t, (p, v), \{u_\tau\}) \cdot v = 0$ for all $t \in \mathbb{N}_0$. It is easy to see that, if $\{u_\tau\}_{\tau \in \mathbb{N}_0} = \{u_{\|_\tau}\}_{\tau \in \mathbb{N}_0} \frac{v}{\|v\|}$, then $\phi_\perp(t, (p, v), \{u_\tau\}) = (p_\perp, 0_d)$ for all $t \in \mathbb{N}_0$. Therefore,

$$\phi(t, (p, v), \{u_\tau\}) = \phi_\|(t, (p, v), \{u_\tau\}) \frac{v}{\|v\|} + (p_\perp, 0_d).$$

Note that, if $p = p_\| \frac{v}{\|v\|} + p_\perp$, then $\|p\| \leq r_{\text{pos}}$ if and only if $p_\| \leq \sqrt{{r_{\text{pos}}}^2 - \|p_\perp\|^2}$. Therefore, the property $\phi\big(t, (p, v), \{u_{\|_\tau}\} \frac{v}{\|v\|}\big) \in \mathcal{A}_0^d(r_{\text{pos}})$ is equivalent to

$$\phi_\| \left(t, (p, v), \{u_{\|_\tau}\} \frac{v}{\|v\|} \right) \in \mathcal{A}_0^1 \Big(\sqrt{r_{\text{pos}}^2 - \|p_\perp\|^2} \Big),$$

and, in turn, definitions (2.7) and (2.8) are equivalent. In order to prove that $\mathcal{A}_\|^d(r_{\text{pos}}, r_{\text{ctr}})$ is compact, we simply observe that it is a closed subset of the compact set $\mathcal{A}^d(r_{\text{pos}}, r_{\text{ctr}})$. □

Remark 2.6. In what follows we use our representation of $\mathcal{A}_\|^d$ to compute an inner approximation for the convex set $\mathcal{A}^d$, for $d \geq 2$. For example, for fixed $p \in B(0_d, r_{\text{pos}})$, we compute velocity vectors v such that $(p, v) \in \mathcal{A}^d$ by considering a sample of unit-length vectors $w \in \mathbb{R}^d$ and computing the maximum allowable velocity v parallel to w according to (2.7). Furthermore, we perform computations by adopting inner polytopic representations for the various compact convex sets. □

2.3. The double-integrator disk graph. Let us introduce some concepts about state-dependent graphs and some useful examples. For a set X, let $\mathbb{F}(X)$ be the collection of finite subsets of X; e.g., $\mathcal{P} \in \mathbb{F}(\mathbb{R}^d)$ is a set of points. For a finite set X, let $\mathbb{G}(X)$ be the set of undirected graphs whose vertices are elements of X, i.e., whose vertex set belongs to $\mathbb{F}(X)$. For a set X, a *state-dependent graph on* X is a map $\mathcal{G} : \mathbb{F}(X) \to \mathbb{G}(X)$ that associates to a finite subset V of X an undirected graph with vertex set V and edge set $\mathcal{E}_{\mathcal{G}}(V)$, where $\mathcal{E}_{\mathcal{G}} : \mathbb{F}(X) \to \mathbb{F}(X \times X)$ satisfies $\mathcal{E}_{\mathcal{G}}(V) \subseteq V \times V$. In other words, what edges exist in $\mathcal{G}(V)$ depends on the elements of V that constitute the nodes.

The following three examples of state-dependent graphs play an important role. First, given $r_{\text{pos}} \in \mathbb{R}_+$, the *disk graph* $\mathcal{G}_{\text{disk}}(r_{\text{pos}})$ is the state-dependent graph on $\mathbb{R}^d$ defined as follows: For $\{p_1, \ldots, p_n\} \subset \mathbb{R}^d$, the pair (p_i, p_j) is an edge in $\mathcal{G}_{\text{disk}}(r_{\text{pos}}) \cdot (\{p_1, \ldots, p_n\})$ if and only if

$$\|p_i - p_j\| \leq r_{\text{pos}} \qquad \Longleftrightarrow \qquad p_i - p_j \in B(0_d, r_{\text{pos}}).$$

Second, given $r_{\text{pos}}, r_{\text{ctr}} \in \mathbb{R}_+$, the *double-integrator disk graph* $\mathcal{G}_{\text{di-disk}}(r_{\text{pos}}, r_{\text{ctr}})$ is the state-dependent graph on $\mathbb{R}^{2d}$ defined as follows: For $\{(p_1, v_1), \ldots, (p_n, v_n)\} \subset \mathbb{R}^{2d}$, the pair $((p_i, v_i), (p_j, v_j))$ is an edge if and only if the relative positions and velocities satisfy

$$(p_i - p_j, v_i - v_j) \in \mathcal{A}^d(r_{\text{pos}}, r_{\text{ctr}}).$$

Third, it is convenient to define the disk graph also as a state-dependent graph on $\mathbb{R}^{2d}$ by stating that $((p_i, v_i), (p_j, v_j))$ is an edge if and only if (p_i, p_j) is an edge of the disk graph on $\mathbb{R}^d$. We illustrate the first two graphs in Figure 2.2.

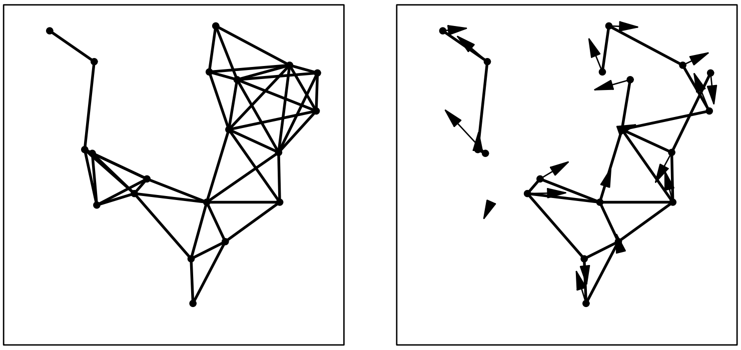


FIG. 2.2. *The disk graph and the double-integrator disk graph in* $\mathbb{R}^2$ *for* 20 *agents with random positions and velocities.*

Remark 2.7. As is well known, the disk graph is invariant under rigid transformations and reflections. Loosely speaking, the double-integrator disk graph is invariant under the following class of transformations: Position and velocities of the agents may be expressed with respect to any rotated and translated frame that is moving at constant linear velocity. These transformations are related to the classic Galilean transformations in geometric mechanics. □

3. Connectivity constraints among second-order agents. In this section we state the model, the notion of connectivity, and a sufficient condition that guarantees connectivity can be preserved.

3.1. Networks of robotic agents with second-order dynamics and the connectivity maintenance problem. We begin by introducing the notion of *network of robotic agents with second-order dynamics* in $\mathbb{R}^d$. Let n be the number of agents. Each agent has the following computation, motion control, and communication capabilities. For $i \in \{1, \dots, n\}$, the ith agent has a processor with the ability of allocating continuous and discrete states and performing operations on them. The ith agent occupies a location $p_i \in \mathbb{R}^d$, moves with velocity $v_i \in \mathbb{R}^d$, according to the discrete time double-integrator dynamics in (2.1), i.e.,

$$\begin{aligned} p_i[t+1] &= p_i[t] + v_i[t], \\ v_i[t+1] &= v_i[t] + u_i[t], \end{aligned} \tag{3.1}$$

where the norm of all controls $u_i[t]$, $i \in \{1, \dots, n\}$, $t \in \mathbb{N}_0$, is upper-bounded by $r_{\text{ctr}} \in \mathbb{R}_+$. The communication model is the following. The processor of each agent has access to the agent's location and velocity. Each agent can transmit information to other agents within a distance $r_{\text{cmm}} \in \mathbb{R}_+$. We remark that the control bound r_{ctr} and the communication radius r_{cmm} are the same for all agents.

Remarks 3.1.

(i) Our network model assumes synchronous execution, although it would be important to consider more general asynchronous networks.

(ii) We will not address in this paper the correctness of our algorithms in the presence of measurement errors or communication quantization.

We now state the control design problem of interest.

PROBLEM 3.2 (connectivity maintenance). *Choose a state-dependent graph $\mathcal{G}_{\text{target}}$ on $\mathbb{R}^{2d}$ and design (state-dependent) control constraints sets with the following property: If each agent's control takes values in the control constraint set, then the agents move in such a way that the number of connected components of $\mathcal{G}_{\text{target}}$ (evaluated at the agents' states) does not increase with time.*

This objective is to be achieved with the limited information available through message exchanges between agents. This problem was originally stated and solved for first-order agents in [2].

3.2. A known result for agents with first-order dynamics. In [2], a connectivity constraint was developed for a set of agents modeled by first-order discrete-time dynamics:

$$p_i[t+1] = p_i[t] + u_i[t].$$

Here the graph whose connectivity is of interest is the disk graph $\mathcal{G}_{\text{disk}}(r_{\text{cmm}})$ over the vertices $\{p_1[t], \dots, p_n[t]\}$. Network connectivity is maintained by restricting the allowable motion of each agent. In particular, it suffices to restrict the motion of each agent as follows. If agents i and j are neighbors in the r_{cmm}-disk graph $\mathcal{G}_{\text{disk}}(r_{\text{cmm}})$ at time t, then their positions at time $t+1$ are required to belong to $B\left(\frac{p_i[t]+p_j[t]}{2}, \frac{r_{\text{cmm}}}{2}\right)$. In other words, connectivity between i and j is maintained if

$$\begin{aligned} u_i[t] &\in B\left(\frac{p_j[t]-p_i[t]}{2}, \frac{r_{\text{cmm}}}{2}\right), \\ u_j[t] &\in B\left(\frac{p_i[t]-p_j[t]}{2}, \frac{r_{\text{cmm}}}{2}\right). \end{aligned}$$

The constraint is illustrated in Figure 3.1.

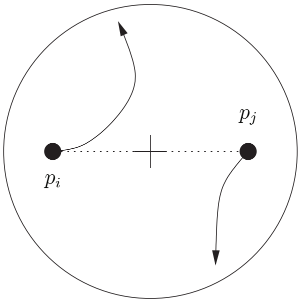


FIG. 3.1. *Starting from p_i and p_j, the agents are restricted to moving inside the disk centered at $\frac{p_i+p_j}{2}$ with radius $\frac{r_{\mathrm{cmm}}}{2}$.*

Note that this constraint takes into account only the positions of the agents; this fact can be attributed to the *first-order dynamics* of the agents. We wish to construct a similar constraint for agents with second-order dynamics. It is reasonable to expect that this new constraint will depend on the positions as well as the velocities of the neighboring agents.

3.3. An appropriate graph for the connectivity maintenance problem for agents with second-order dynamics. We begin working on the stated problem with a negative result regarding two candidate target graphs.

LEMMA 3.3 (unsuitable graphs). *Consider a network of n agents with double integrator dynamics* (3.1) *in $\mathbb{R}^d$. Let r_{cmm} be the communication range and let r_{ctr} be the control bound. Let $\mathcal{G}_{\mathrm{target}}$ be either $\mathcal{G}_{\mathrm{disk}}(r_{\mathrm{cmm}})$ on $\mathbb{R}^{2d}$ or $\mathcal{G}_{\text{di-disk}}(r_{\mathrm{cmm}}, 2r_{\mathrm{ctr}})$. There exist states $\{(p_i, v_i)\}_{i\in\{1,\dots,n\}}$ such that*

(i) *the graph $\mathcal{G}_{\mathrm{target}}$ at $\{(p_i, v_i)\}_{i\in\{1,\dots,n\}}$ is connected, and*

(ii) *for all $\{u_i\}_{i\in\{1,\dots,n\}} \subseteq B(0_d, r_{\mathrm{ctr}})$, the graph $\mathcal{G}_{\mathrm{target}}$ at $\{(p_i+v_i, v_i+u_i)\}_{i\in\{1,\dots,n\}}$ is disconnected.*

Proof. The proof of the statement for $\mathcal{G}_{\mathrm{target}} = \mathcal{G}_{\mathrm{disk}}(r_{\mathrm{cmm}})$ is straightforward. Consider two agents whose relative position and velocity belong to $\mathcal{A}_0^d \setminus \mathcal{A}_1^d$. Then, after one time step, the two agents will not be connected in $\mathcal{G}_{\mathrm{disk}}(r_{\mathrm{cmm}})$ no matter what their controls are. Next, consider the case $\mathcal{G}_{\mathrm{target}} = \mathcal{G}_{\text{di-disk}}(r_{\mathrm{cmm}}, 2r_{\mathrm{ctr}})$. For $d = 1$, let $\bar{v}$ be the maximal velocity in $\mathcal{A}^1(r_{\mathrm{cmm}}, 2r_{\mathrm{ctr}})$ at $p = 0$, that is, $\bar{v} = \min\{r_{\mathrm{cmm}}/m + (m-1)r_{\mathrm{ctr}} \mid m \in \mathbb{N}\}$. Take three agents with positions $p_1 = p_2 = p_3 = 0$ and velocities $v_1 = -\bar{v}$, $v_2 = 0$, and $v_3 = \bar{v}$. At this configuration, the graph $\mathcal{G}_{\text{di-disk}}(r_{\mathrm{cmm}}, 2r_{\mathrm{ctr}})$ contains two edges: one between agents 1 and 2, and the other between agents 2 and 3. Connectivity is preserved after one time step if agent 2 remains connected to both agents 1 and 3 after one time step. However, to remain connected to agent 1, its control u_2 must be equal to $-r_{\mathrm{ctr}}$ and, analogously, to remain connected with agent 3, its control u_2 must be equal to $+r_{\mathrm{ctr}}$. Clearly this is impossible. □

Remarks 3.4.

(i) The result in Lemma 3.3 on the double integrator graph has the following interpretation. Assume that agent i has two neighbors j and k in the graph $\mathcal{G}_{\text{di-disk}}(r_{\mathrm{cmm}}, r_{\mathrm{ctr}})$. By definition of the neighboring law for this graph, we know that there exists bounded controls for i and j to maintain the $((p_i, v_i), (p_j, v_j))$ link and that there exists bounded controls for i and k to maintain the $((p_i, v_i), (p_k, v_k))$ link. The lemma states that, for some states of the agents i, j, and k, there might not exist controls that maintain both links simultaneously.

(ii) In other words, Lemma 3.3 shows how the disk graph $\mathcal{G}_{\text{disk}}(r_{\text{cmm}})$ and the double integrator disk graph $\mathcal{G}_{\text{di-disk}}(r_{\text{cmm}}, 2r_{\text{ctr}})$ are not appropriate choices for the connectivity maintenance problem.

The following theorem suggests that an appropriate scaling of the control bound is helpful in identifying a suitable state-dependent graph for Problem 3.2.

THEOREM 3.5 (a scaled double-integrator disk graph is suitable). *Consider a network of n agents with double-integrator dynamics* (3.1) *in $\mathbb{R}^d$. Let r_{cmm} be the communication range and let r_{ctr} be the control bound. For $k \in \{1, \dots, n-1\}$, define*

$$\nu(k) = \frac{2}{k\sqrt{d}}.$$

Assume that $k \in \{1, \dots, n-1\}$ and the state $\{(p_i, v_i)\}_{i\in\{1,\dots,n\}}$ together have the property that the graph $\mathcal{G}_{\text{di-disk}}(r_{\text{cmm}}, \nu(k)r_{\text{ctr}})$ at $\{(p_i, v_i)\}_{i\in\{1,\dots,n\}}$ contains a spanning tree T with diameter at most k. Then there exists $\{u_i\}_{i\in\{1,\dots,n\}} \subseteq B(0_d, r_{\text{ctr}})$ such that if $((p_i, v_i), (p_j, v_j))$ is an edge of T, then $((p_i + v_i, v_i + u_i), (p_j + v_j, v_j + u_j))$ is an edge of $\mathcal{G}_{\text{di-disk}}(r_{\text{cmm}}, \nu(k)r_{\text{ctr}})$ at $\{(p_i + v_i, v_i + u_i)\}_{i\in\{1,\dots,n\}}$.

These results are based upon Shostak's theory for systems of inequalities, as exposed in [10]. We provide the proof in the appendix. The following results are immediate consequences of this theorem.

COROLLARY 3.6 (simple sufficient condition). *With the same notation in Theorem* 3.5*, define $\nu_{\min} = \frac{2}{(n-1)\sqrt{d}}$. Assume that the state $\{(p_i, v_i)\}_{i\in\{1,\dots,n\}}$ has the property that the graph $\mathcal{G}_{\text{di-disk}}(r_{\text{cmm}}, \nu_{\min}r_{\text{ctr}})$ is connected at $\{(p_i, v_i)\}_{i\in\{1,\dots,n\}}$. Then*

(i) *there exists $\{u_i\}_{i\in\{1,\dots,n\}} \subseteq B(0_d, r_{\text{ctr}})$, such that the graph $\mathcal{G}_{\text{di-disk}}(r_{\text{cmm}}, \nu_{\min}r_{\text{ctr}})$ is also connected at $\{(p_i + v_i, v_i + u_i)\}_{i\in\{1,\dots,n\}}$; and*

(ii) *if T is a spanning tree of $\mathcal{G}_{\text{di-disk}}(r_{\text{cmm}}, \nu_{\min}r_{\text{ctr}})$ at $\{(p_i, v_i)\}_{i\in\{1,\dots,n\}}$, then there exists $\{u_i\}_{i\in\{1,\dots,n\}} \subseteq B(0_d, r_{\text{ctr}})$ such that for all edges $((p_i, v_i), (p_j, v_j))$ of T, it holds that $((p_i + v_i, v_i + u_i), (p_j + v_j, v_j + u_j))$ is an edge of $\mathcal{G}_{\text{di-disk}}(r_{\text{cmm}}, \nu_{\min}r_{\text{ctr}})$ at $\{(p_i + v_i, v_i + u_i)\}_{i\in\{1,\dots,n\}}$.*

Remark 3.7 (scaling of $\nu_{\min}$ with n). The condition $\nu_{\min} = \frac{2}{\sqrt{d}(n-1)}$ implies that, according to the sufficient conditions in Corollary 3.6, as the number of agents grows, the velocities of the agents must be closer and closer in order for the agents to be able to remain connected.

If $\mathcal{G}_{\text{di-disk}}(r_{\text{cmm}}, \nu(k)r_{\text{ctr}})$ at $\{(p_i, v_i)\}_{i\in\{1,\dots,n\}}$ is not connected for some k, then Theorem 3.5 applies to its connected components. In what follows we fix k and without loss of generality assume the graph $\mathcal{G}_{\text{di-disk}}(r_{\text{cmm}}, \nu(k)r_{\text{ctr}})$ at $\{(p_i, v_i)\}_{i\in\{1,\dots,n\}}$ to be connected. □

Remark 3.8 (distributed computation of connectivity and of spanning trees). Each agent can compute its neighbors in the graph $\mathcal{G}_{\text{di-disk}}(r_{\text{cmm}}, \nu(k)r_{\text{ctr}})$ just by communicating with its neighbors in $\mathcal{G}_{\text{disk}}(r_{\text{cmm}})$ and exchanging with them position and velocity information. Alternatively, this computation may also be performed if each agent may sense relative position and velocity with all other agents at a distance no larger than r_{cmm}.

It is possible to compute spanning trees in networks via standard depth-first search distributed algorithms. It is also possible [11] to distributively compute a minimum diameter spanning tree in a network. □

3.4. The control constraint set and its polytopic representation. We now concentrate on two agents with indices i and j. For $t \in \mathbb{N}_0$, we define the relative position, velocity, and control by $p_{ij}[t] = p_i[t] - p_j[t]$, $v_{ij}[t] = v_i[t] - v_j[t]$, and $u_{ij}[t] = u_i[t] - u_j[t]$, respectively. It is easy to see that

$$\begin{aligned} p_{ij}[t+1] &= p_{ij}[t] + v_{ij}[t], \\ v_{ij}[t+1] &= v_{ij}[t] + u_{ij}[t]. \end{aligned}$$

Assume that agents i, j are connected in $\mathcal{G}_{\text{di-disk}}(r_{\text{cmm}}, \nu(k) r_{\text{ctr}})$ at time t. By definition, this means that the relative state $(p_{ij}[t], v_{ij}[t])$ belongs to $\mathcal{A}^d(r_{\text{cmm}}, \nu(k) r_{\text{ctr}})$. If this connection is to be maintained at time $t+1$, then the relative control at time t must satisfy

$$u_i[t] - u_j[t] \in \mathcal{U}^d(r_{\text{cmm}}, \nu(k) r_{\text{ctr}}) \cdot (p_{ij}[t], v_{ij}[t]). \tag{3.2}$$

Also, implicit are the following bounds on individual control inputs $u_i[t]$ and $u_j[t]$:

$$u_i[t] \in B(0_d, r_{\text{ctr}}), \quad u_j[t] \in B(0_d, r_{\text{ctr}}). \tag{3.3}$$

This discussion motivates the following definition.

Definition 3.9. *Given* $r_{\text{cmm}}, r_{\text{ctr}}, \nu(k) \in \mathbb{R}_+$ *and given a set* E *of edges in* $\mathcal{G}_{\text{di-disk}}(r_{\text{cmm}}, \nu(k) r_{\text{ctr}})$ *at* $\{(p_i, v_i)\}_{i \in \{1,\dots,n\}}$, *the* control constraint set *is defined by*

$$\begin{aligned} &\mathcal{U}^d_E(r_{\text{cmm}}, r_{\text{ctr}}, \nu(k)) \cdot (\{p_i, v_i\}_{i \in \{1,\dots,n\}}) \\ &\qquad = \{(u_1, \dots, u_n) \in B(0_d, r_{\text{ctr}})^n \mid \forall ((p_i, v_i), (p_j, v_j)) \in E, \\ &\qquad\qquad u_i - u_j \in \mathcal{U}^d(r_{\text{cmm}}, \nu(k) r_{\text{ctr}}) \cdot (p_i - p_j, v_i - v_j)\}. \end{aligned}$$

In other words, the control constraint set for an edge set E is the set of controls for each agent with the property that all edges in E will be maintained in one time step.

Remark 3.10. We can now interpret the results in Theorem 3.5 as follows.

(i) To maintain connectivity between any pair of connected agents, we should simultaneously handle constraints corresponding to *all* edges of $\mathcal{G}_{\text{di-disk}}(r_{\text{cmm}}, \nu(k) r_{\text{ctr}})$. This might render the control constraint set empty.

(ii) However, if we only consider constraints corresponding to edges belonging to *a* spanning tree T of $\mathcal{G}_{\text{di-disk}}(r_{\text{cmm}}, \nu(k) r_{\text{ctr}})$, then the set $\mathcal{U}^d_T(r_{\text{cmm}}, \nu(k) r_{\text{ctr}}) \cdot (\{p_i, v_i\}_{i \in \{1,\dots,n\}})$ is guaranteed to be nonempty.

Let us now provide a concrete representation of the control constraint set. Given a pair i, j of connected agents, the admissible control set $\mathcal{U}^d(r_{\text{cmm}}, \nu(k) r_{\text{ctr}}) \cdot (p_{ij}, v_{ij})$ is convex and compact (Lemma 2.2). Hence, we can fit a polytope with N_{poly} sides inside it. This approximating polytope leads to the following tighter version of the constraint in (3.2):

$$(C^\eta_{ij})^T (u_i - u_j) \le w^\eta_{ij}, \qquad \eta \in \{1, \dots, N_{\text{poly}}\}, \tag{3.4}$$

for some appropriate vector $C^\eta_{ij} \in \mathbb{R}^d$ and scalar $w^\eta_{ij} \in \mathbb{R}$. Similarly, one can compute an inner polytopic approximation of the closed ball $B(0_d, r_{\text{ctr}})$ and write the following linear vector inequalities:

$$(C^\eta_{i\theta})^T u_i \le w^\eta_{i\theta}, \qquad \eta \in \{1, \dots, N_{\text{poly}}\}, \tag{3.5}$$

where the symbol θ has the interpretation of a fictional agent. In this way, we have cast the original problem of finding a set of feasible control inputs into a satisfiability problem for a set of linear inequalities.

Remark 3.11. Rather than a network-wide control constraint set, one might like to obtain decoupled constraint sets for each individual agent. However, (1) it is not clear how to design a distributed algorithm to perform this computation, (2) such an algorithm will likely have large communication requirements, and (3) such a calculation might lead to a very conservative estimate for the decoupled control constraint sets. Therefore, rather than explicitly decoupling the control constraint sets, we next focus on a distributed algorithm to search the control constraint set for feasible controls that are optimal according to some criterion. □

4. Distributed computation of optimal controls.

In the previous section, we derived sufficient conditions for the existence of connectivity-maintaining control inputs. In this section, we utilize these analysis results to tackle a design problem. We provide an algorithm to compute connectivity-maintaining control inputs and we do so satisfying two requirements: We require the algorithm to be distributed and to be optimal in the following sense. We assume a "high level" controller is available to compute desired control inputs for the agents to achieve a specific task independent of the connectivity-maintaining constraints. We then design a "low level" filter which computes, inside the set of connectivity-maintaining inputs, the closest inputs to the desired ones. We set up this filter design problem in the form of an optimization problem with the performance criteria being the minimization of the (squared) Euclidean norm of the deviation away from the desired inputs. The resulting quadratic optimization problem can be solved through a distributed "flow control" algorithm. As an example application, we illustrate by simulations that solving this optimization problem for a simple choice of the desired inputs achieves a weak form of connectivity-preserving "flocking" behavior among the agents.

4.1. Problem formulation.

We consider the following optimization problem: Given an array of desired control inputs $U_{\text{des}} = (u_{\text{des},1}, \dots, u_{\text{des},n})^T \in (\mathbb{R}^d)^n$, find, via local computation, the array $U = (u_1, \dots, u_n)$ belonging to the control constraint set, that is *closest* to the desired array U_{des}. To formulate this problem precisely, we need some additional notations. Let E be a set of edges in the undirected graph $\mathcal{G}_{\text{di-disk}}(r_{\text{cmm}}, \nu(k) r_{\text{ctr}})$ at $\{(p_i, v_i)\}_{i \in \{1,\dots,n\}}$. To deal with the linear inequalities of the forms (3.4) and (3.5) associated to each edge of E, we introduce an appropriate *multigraph*. A *multigraph* (or *multiple edge graph*) is, roughly speaking, a graph with multiple edges between the same vertices. More formally, a multigraph is a pair $(V_{\text{mult}}, E_{\text{mult}})$, where V_{mult} is the vertex set and the edge set E_{mult} contains numbered edges of the form (i, j, η), for $i, j \in V$ and $\eta \in \mathbb{N}$, and where E_{mult} has the property that if $(i, j, \eta) \in E_{\text{mult}}$ and $\eta > 1$, then also $(i, j, \eta - 1) \in E_{\text{mult}}$. In what follows, we let G_{mult} denote the multigraph with vertex set $\{1, \dots, n\}$ and with edge set $E_{\text{mult}} = \{(i, j, \eta) \in \{1, \dots, n\}^2 \times \{1, \dots, N_{\text{poly}}\} \mid ((p_i, v_i), (p_j, v_j)) \in E,\ i > j\}$. Note that to each element $(i, j, \eta) \in E_{\text{mult}}$ is associated the inequality $(C_{ij}^{\eta})^T (u_i - u_j) \le w_{ij}^{\eta}$. We are now ready to formally state the optimization problem at hand in the form of the following quadratic programming problem:

$$
\begin{aligned}
&\text{Minimize} \quad && \frac{1}{2} \sum_{i=1}^{n} \|u_i - u_{\text{des},i}\|^2 \\
&\quad\text{s.t.} && (C_{ij}^{\eta})^T (u_i - u_j) \le w_{ij}^{\eta} \quad \text{for } (i, j, \eta) \in E_{\text{mult}}, \\
& && (C_{i\theta}^{\eta})^T u_i \le w_{i\theta}^{\eta} \qquad \text{for } i \in \{1, \dots, n\}, \eta \in \{1, \dots, N_{\text{poly}}\}.
\end{aligned}
\tag{4.1}
$$

Here, somehow arbitrarily, we have adopted the 2-norm to define the cost function.

Remark 4.1 (feasibility). If E is a spanning tree of $\mathcal{G}_{\text{di-disk}}(r_{\text{cmm}}, \nu r_{\text{ctr}})$ at a connected configuration $\{(p_i, v_i)\}_{i\in\{1,\dots,n\}}$, then the control constraint set $\mathcal{U}^d_E(r_{\text{cmm}}, r_{\text{ctr}}, \nu(k))\cdot(\{p_i, v_i\}_{i\in\{1,\dots,n\}})$ is guaranteed to be nonempty by Theorem 3.5. In turn, this implies that the optimization problem (4.1) is feasible. □

4.2. Solution via duality: The projected Jacobi method. The problem (4.1) can be written in a compact form as

$$\begin{aligned} \text{minimize } & \frac{1}{2}\|U - U_{\text{des}}\|^2 \\ \text{s.t. } & B^T_{\text{mult}} U \preceq w, \end{aligned}$$

for appropriately defined matrix B_{mult} and vector w. A dual "projected Jacobi method" algorithm for the solution of this standard quadratic program is described in [12]. According to this algorithm, let λ^* be the value of Lagrange multipliers at optimality. Then the global minimum for U is achieved at

$$U^* = U_{\text{des}} - B_{\text{mult}}\lambda^*. \tag{4.2}$$

The projected Jacobi iteration for each component of λ is given by

$$\begin{aligned} \lambda_\alpha(t+1) = \max\Bigg\{0, \lambda_\alpha(t) - \frac{\tau}{(B^T_{\text{mult}}B_{\text{mult}})_{\alpha\alpha}}\Bigg((w - B^T_{\text{mult}}U_{\text{des}})_\alpha \\ + \sum_{\beta=1}^{N_{\text{poly}}(e+n)} (B^T_{\text{mult}}B_{\text{mult}})_{\alpha\beta}\lambda_\beta(t)\Bigg)\Bigg\}, \end{aligned} \tag{4.3}$$

where $\alpha \in \{1,\dots,N_{\text{poly}}(e+n)\}$ and τ is the step size parameter. We can select $\tau = \frac{1}{N_{\text{poly}}(e+n)}$ to guarantee convergence.

4.3. A distributed implementation of the dual algorithm. Because of the particular structure of the matrix $B^T_{\text{mult}}B_{\text{mult}}$, the iterations (4.3) can be implemented in a distributed way over the original graph G. To highlight the distributed structure of the iteration we denote the components of λ by referring to the nodes that they share and the inequality to which they are related. In particular for each edge in G_{mult}, the corresponding Lagrange multiplier will be denoted as λ^η_{ij} if the edge goes from node i to node j, $i > j$, and the edge is associated to the inequality constraint $C^\eta_{ij}(u_i - u_j) \leq w^\eta_{ij}$. This makes up the first $N_{\text{poly}}e$ entries of the vector λ. To be consistent with this notation, the next $N_{\text{poly}}n$ entries will be denoted $\lambda^1_{1\theta}, \dots, \lambda^{N_{\text{poly}}}_{1\theta}, \dots, \lambda^1_{n\theta}, \dots, \lambda^{N_{\text{poly}}}_{n\theta}$. Additionally, define $\mathcal{N}(i) = \{j \in \{1,\dots,n\} \mid \{(p_i, v_i), (p_j, v_j)\} \in E\} \cup \{\theta\}$. The symbol θ has the interpretation of a fictional node.

Defining $\lambda^\eta_{ij} := \lambda^\eta_{ji}$ and $C^\eta_{ij} := -C^\eta_{ji}$ for $i < j$, we can write (4.2) and (4.3) in components as follows. Equation (4.2) reads, for $i \in \{1,\dots,n\}$,

$$u^*_i = u_{\text{des},i} - \sum_{k\in\mathcal{N}(i)} \sum_{\eta=1}^{N_{\text{poly}}} C^\eta_{ik}\lambda^\eta_{ik}.$$

One can easily work out an explicit expression for matrix product $B^T_{\text{mult}}B_{\text{mult}}$ in (4.3).

Then, (4.3) reads, for $(i, j, \eta) \in E_{\text{mult}}$,

$$\lambda_{ij}^{\eta}(t+1) = \max\Bigg\{0, \lambda_{ij}^{\eta}(t) - \frac{\tau}{2(C_{ij}^{\eta})^T C_{ij}^{\eta}} \cdot \Bigg(\sum_{k \in \mathcal{N}(i)} \sum_{\sigma=1}^{N_{\text{poly}}} \Big((C_{ij}^{\eta})^T C_{ik}^{\sigma} \lambda_{ik}^{\sigma} \Big) + \sum_{k \in \mathcal{N}(j)} \sum_{\sigma=1}^{N_{\text{poly}}} \Big((C_{ji}^{\eta})^T C_{jk}^{\sigma} \lambda_{jk}^{\sigma} \Big) + w_{ij}^{\eta} - (C_{ij}^{\eta})^T (u_{\text{des},i} - u_{\text{des},j}) \Bigg) \Bigg\},$$

together with, for $i \in \{1, \dots, n\}$, $\eta \in \{1, \dots, N_{\text{poly}}\}$,

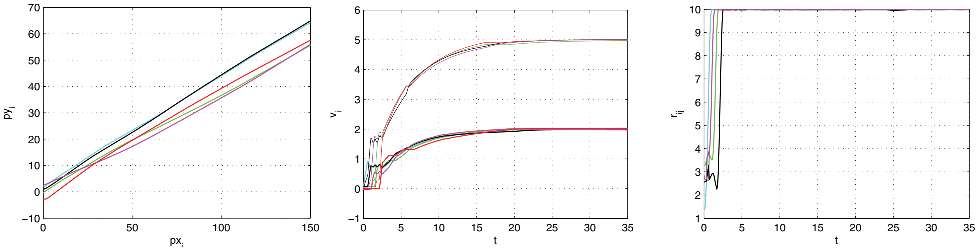
$$\lambda_{i\theta}^{\eta}(t+1) = \max\Bigg\{0, \lambda_{i\theta}^{\eta}(t) - \frac{\tau}{(C_{i\theta}^{\eta})^T C_{i\theta}^{\eta}} \Bigg(\sum_{k \in \mathcal{N}(i)} \sum_{\sigma=1}^{N_{\text{poly}}} ((C_{i\theta}^{\eta})^T C_{ik}^{\sigma} \lambda_{ik}^{\sigma}) + w_{i\theta}^{\eta} - (C_{i\theta}^{\eta})^T u_{\text{des},i} \Bigg) \Bigg\}.$$

We distribute the task of running iterations for these $N_{\text{poly}}(e+n)$ Lagrange multipliers among the n agents as follows: An agent i carries out the updates for all quantities $\lambda_{i\theta}^{\eta}$ and all λ_{ij}^{η} for which $i > j$. By means of this partition and by means of iterated one-hop communication among agents, it is possible to compute the global solution for the optimization problem (4.1) in a distributed fashion over the double integrator disk graph.

4.4. Simulations. To illustrate our analysis we focus on the following scenario. For the two-dimensional setting, i.e., for $d = 2$, we assume that there are $n = 5$ agents with (randomly chosen) initial condition and such that they are connected according to $\mathcal{G}_{\text{di-disk}}$. The bound for the control input is $r_{\text{ctr}} = 2$ and the communication radius is $r_{\text{cmm}} = 10$.

We assigned to one of the agents a derivative feedback control $u_x[p, v] = (v_x - 2)$, $u_y[p, v] = (v_y - 5)$ as desired input. For the other agents the desired input is set to zero. We show the agent trajectories in Figure 4.1(a), the velocities of the agents with respect to time in Figure 4.1(b), and the distances between agents which are neighbors in the spanning tree in Figure 4.1(c). Notice that the agents move with approximately identical velocity reaching a configuration in which all of them are at the limit distance $r_{\text{cmm}} = 10$. The interesting aspect of this simulation is that the maintenance of connectivity leads to the accomplishment of apparently unrelated coordination tasks as velocity alignment and cohesiveness. This numerical result illustrate how our connectivity maintenance approach might indeed be a starting point for novel investigations into the problem of flocking with connectivity.

5. Conclusion. We provide some distributed algorithms to enforce connectivity among networks of agents with double-integrator dynamics. Future directions of research include (i) evaluating the communication complexity of the proposed distributed dual algorithm and possibly designing faster ones, (ii) studying the relationship between the connectivity maintenance problem and the platooning and mesh



(a) Positions (b) Velocities (v_x and v_y) (c) Interagent distances

FIG. 4.1. *Velocity alignment and cohesiveness for five agents in the plane ($d = 2$).*

stability problem, and (iii) investigating the flocking phenomenon and designing flocking algorithms which do not rely on a blanket assumption of connectivity.

Appendix. Shostak's test. This section provides a proof for Theorem 3.5. The proof amounts to showing that if E is the edge set of a spanning tree T in $\mathcal{G}_{\text{di-disk}}(r_{\text{cmm}}, \nu(k) r_{\text{ctr}})$ at $\{(p_i, v_i)\}_{i \in \{1,\dots,n\}}$, then the control constraint set $\mathcal{U}^d_E(r_{\text{cmm}}, r_{\text{ctr}}, \nu(k)) \cdot (\{p_i, v_i\}_{i \in \{1,\dots,n\}})$ is nonempty. We first consider a polytopic approximation of constraints (3.2) and (3.3). Among all possible choices, we use the conservative orthotope approximation that allows us to decouple the constraints into d independent sets of linear inequalities (one for each dimension). Then we use Shostak's theory to obtain sufficient conditions for the feasibility of these linear inequalities. For brevity, we drop the dependence of the quantities on t and we assume that the variables u_i are scalars for all $i \in \{1, \dots, n\}$ and $t \geq 0$. The resulting sets of linear inequalities for one particular dimension are

$$\delta^l_{i,j} \leq u_i - u_j \leq \delta^u_{i,j} \quad \text{and} \quad -\frac{r_{\text{ctr}}}{\sqrt{d}} \leq u_i \leq \frac{r_{\text{ctr}}}{\sqrt{d}}, \tag{A.1}$$

where $-\nu(k) r_{\text{ctr}} \leq \delta^l_{i,j} \leq \delta^u_{i,j} \leq \nu(k) r_{\text{ctr}}$ for all $i, j \in \{1, \dots, n\}$ and $i \neq j$.

A.1. Shostak's theory. In this section we present Shostak's theory for feasibility of linear inequalities involving at most two variables, similar to the ones in (A.1). These ideas will then be used to prove Theorem 3.5. The notations used in [10] adapted to our case are presented next. Let u_0 be an auxiliary *zero variable* that always occurs with zero coefficient—the only variable that can do this. Without loss of generality, we can thus assume that all the inequalities in $\mathcal{L}$ contain two variables. As a result of this, the inequalities in (A.1) can be succinctly written as

$$u_i - u_j \leq \delta_{i,j} \quad \forall i, j \in \{0, \dots, n\}, \tag{A.2}$$

where for all $i, j \in \{1, \dots, n\}$, $i \neq j, -\nu(k) r_{\text{ctr}} \leq \delta_{i,j} \leq \nu(k) r_{\text{ctr}}$ and for all $i \in \{1, \dots, n\}, \delta_{i,0} = \delta_{0,i} = \frac{r_{\text{ctr}}}{\sqrt{d}}$. Also implicit in this formulation is the relation that $\delta_{i,j} + \delta_{j,i} \geq 0$ for all $i, j \in \{0, \dots, n\}$ and $i \neq j$.

Let $\mathcal{L}$ denote the system of inequalities in (A.2). We construct the graph $G(\mathcal{L})$ with $n + 1$ vertices and $2(2n - 1)$ edges as follows: (a) For each variable u_i occurring in $\mathcal{L}$, add a vertex i to $G(\mathcal{L})$. (b) For each inequality of the form $a_{i,j} u_i + b_{i,j} u_j \leq \delta_{i,j}$ in $\mathcal{L}$, add an undirected edge between i and j to $G(\mathcal{L})$, and label the edge with the inequality (see Figure A.1). It is easy to see the following relations between the spanning tree T in $\mathcal{G}_{\text{di-disk}}(r_{\text{cmm}}, \nu(k) r_{\text{ctr}})$ at $\{(p_i, v_i)\}_{i \in \{1,\dots,n\}}$ that is used to derive

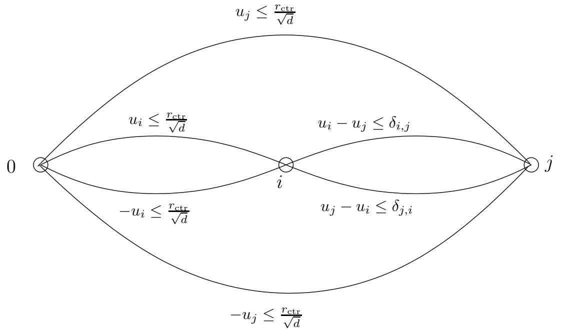

$u_j \leq \frac{r_{\mathrm{ctr}}}{\sqrt{d}}$
$u_i \leq \frac{r_{\mathrm{ctr}}}{\sqrt{d}}$
$u_i - u_j \leq \delta_{i,j}$
0
i
j
$-u_i \leq \frac{r_{\mathrm{ctr}}}{\sqrt{d}}$
$u_j - u_i \leq \delta_{j,i}$
$-u_j \leq \frac{r_{\mathrm{ctr}}}{\sqrt{d}}$

FIG. A.1. *Snippet of the graph $G(\mathcal{L})$ for the system of inequalities in* (A.2).

the constraints in the inequalities (A.2) and the graph $G(\mathcal{L})$: (a) The vertex set of $G(\mathcal{L})$ is the union of the vertex set of T and the auxiliary vertex 0; (b) for every edge $\{i, j\}$ in T, there are two edges between the vertices i and j in $G(\mathcal{L})$; (c) additionally, $G(\mathcal{L})$ contains two edges between 0 and every other vertex i for all $i \in \{1, \dots, n\}$.

To every edge represented by the inequality of the form $a_{i,j}u_i + b_{i,j}u_j \leq \delta_{i,j}$, we associate *a triple* $\langle a_{i,j}, b_{i,j}, \delta_{i,j}\rangle$. Note that $\langle b_{i,j}, a_{i,j}, \delta_{i,j}\rangle$ is also a triple associated with the same edge. Without loss of generality, consider a path of $G(\mathcal{L})$ determined by the vertices $\{1, 2, \dots, l+1\}$ and the edges $e_{1,2}, e_{2,3}, \dots, e_{l,l+1}$ between them. *A triple sequence*, P, associated with the path is defined as

$$\langle a_{1,2}, b_{1,2}, \delta_{1,2}\rangle, \langle a_{2,3}, b_{2,3}, \delta_{2,3}\rangle, \dots, \langle a_{l,l+1}, b_{l,l+1}, \delta_{l,l+1}\rangle,$$

where, for $1 \leq i \leq l$, $a_{i,i+1}u_i + b_{i,i+1}u_j \leq \delta_{i,i+1}$ is the inequality associated with the edge $e_{i,i+1}$. If $a_{i+1,i+2}$ and $b_{i,i+1}$ have opposite signs for $1 \leq i < l$, then P is called *admissible.*

Define $\langle a_P, b_P, \delta_P\rangle$, the *residue* of P, as

$$\langle a_P, b_P, \delta_P\rangle = \langle a_{1,2}, b_{1,2}, \delta_{1,2}\rangle \odot \langle a_{2,3}, b_{2,3}, \delta_{2,3}\rangle \odot \cdots \odot \langle a_{l,l+1}, b_{l,l+1}, \delta_{l,l+1}\rangle,$$

where $\odot$ is the associativity binary operator defined on triples by

$$\langle a, b, \delta\rangle \odot \langle a', b', \delta'\rangle = \langle \kappa a a', -\kappa b b', \kappa(\delta a' - \delta' b)\rangle, \\ \text{where} \quad \kappa = a'/|a'|.$$

Intuitively, the operator $\odot$ takes two inequalities and derives a new inequality by eliminating a common variable; e.g., $ax + by \leq \delta$ and $a'y + b'z \leq \delta'$ imply $-aa'x + bb'z \leq -(\delta a' - \delta' b)$ if $a < 0$ and $b > 0$. Note that the signs of a_P and $a_{1,2}$ agree, as do the signs of b_P and $b_{1,2}$.

A path is called a *loop* if the initial and final vertices are identical. (A loop is not uniquely specified unless its initial vertex is given.) If all the intermediate vertices of a path are distinct, the path is *simple.* An admissible triple sequence P associated with a loop with initial vertex x is *infeasible* if its residue satisfies $a_P + b_P = 0$ and $\delta_P < 0$. A loop which contains an infeasible triple sequence is called an *infeasible loop.* Thus if $G(\mathcal{L})$ has an infeasible loop, the system of inequalities $\mathcal{L}$ is unsatisfiable. However, the converse is not true in general. Next, we show how to extend $\mathcal{L}$ to an equivalent system $\mathcal{L}'$ such that $G(\mathcal{L}')$ has an infeasible simple loop if and only if $\mathcal{L}$ is unsatisfiable.

For each vertex i of $G(\mathcal{L})$ and for each admissible triple sequence P with $a_P+b_P \neq 0$ associated with a simple loop of $G(\mathcal{L})$ and initial vertex i, add a new inequality $(a_P+b_P)u_i \leq \delta_P$ to $\mathcal{L}$. This new system $\mathcal{L}'$ is referred to as the *Shostak extension* of $\mathcal{L}$. We now state the necessary and sufficient condition on the extended system of inequalities $\mathcal{L}'$ for the satisfiability of the original system $\mathcal{L}$.

THEOREM A.1 (see Shostak's theorem [10]). *Let $\mathcal{L}'$ be the Shostak extension of $\mathcal{L}$. The system of inequalities $\mathcal{L}$ is satisfiable if and only if $G(\mathcal{L}')$ contains no infeasible simple loop.*

A.2. Satisfiability test. In this section we use the Shostak criterion to derive conditions for the satisfiability of the inequalities in (A.2).

LEMMA A.2. *Let $\mathcal{L}$ be the system of inequalities of the form* (A.2) *obtained by considering pairwise neighbors in a spanning tree T in $\mathcal{G}_{\text{di-disk}}(r_{\text{cmm}}, \nu(k)r_{\text{ctr}})$ at $\{(p_i,v_i)\}_{i\in\{1,\dots,n\}}$. Then the Shostak extension of $\mathcal{L}$ is itself.*

Proof. Consider a simple loop of $G(\mathcal{L})$ with the initial vertex $i \in \{0,1,\dots,n\}$. Consider an admissible triple sequence P associated with the loop. Since $a_{i,j}, b_{i,j} \in \{-1,+1\}$ for all $i,j \in \{1,\dots,n\}, i \neq j$, and $a_{0,i}, a_{i,0}, b_{i,0}, b_{0,i} \in \{-1,0,+1\}$ for all $i \in \{1,\dots,n\}$, the residue of P, $\langle a_P, b_P, \delta_P\rangle$ is such that $a_p + b_p = 0$. Hence, no new inequality must be added to obtain the Shostak extension of $\mathcal{L}$. □

LEMMA A.3. *Let $\mathcal{L}$ be the system of inequalities of the form* (A.2) *obtained by considering pairwise neighbors in a spanning tree T of depth at most k in $\mathcal{G}_{\text{di-disk}}(r_{\text{cmm}}, \nu(k)r_{\text{ctr}})$ at $\{(p_i,v_i)\}_{i\in\{1,\dots,n\}}$. If $\nu(k) = \frac{2}{k\sqrt{d}}$, then there is no infeasible simple loop in $G(\mathcal{L})$.*

Proof. Looking at Figure A.1, it is clear that there are two types of simple loops with admissible triple sequences in $G(\mathcal{L})$:

(i) $\langle +1,-1,\delta_{i,j}\rangle$, $\langle +1,-1,\delta_{j,i}\rangle$, or $\langle -1,+1,\delta_{i,j}\rangle$, $\langle -1,+1,\delta_{j,i}\rangle$, where $i,j \in \{0,\dots,n-1\}$ and $\{i,j\}$ is an edge in T.

(ii) $\langle 0,-1,\frac{r_{\text{ctr}}}{\sqrt{d}}\rangle, \langle +1,-1,\delta_{i_1,i_2}\rangle, \dots, \langle +1,-1,\delta_{i_{l-1},i_l}\rangle, \langle +1,0,\frac{r_{\text{ctr}}}{\sqrt{d}}\rangle$ or $\langle 0,+1,\frac{r_{\text{ctr}}}{\sqrt{d}}\rangle, \langle -1,+1,\delta_{i_2,i_1}\rangle, \dots, \langle -1,+1,\delta_{i_l,i_{l,l-1}}\rangle, \langle -1,0,\frac{r_{\text{ctr}}}{\sqrt{d}}\rangle$, where $i_l \in \{1,\dots,\zeta\}$ for all $l \in \{1,\dots,\zeta\}$ and $\{i_l, i_{l+1}\}$ is an edge in T.

The residue for the first set of loops is $\langle +1,-1,\delta_{i,j}+\delta_{j,i}\rangle$ or $\langle -1,+1,\delta_{i,j}+\delta_{j,i}\rangle$. The feasibility condition is trivially satisfied by construction since $\delta_{i,j}+\delta_{j,i} \geq 0$. For the second set of loops, the residue is

$$\left\langle 0,-1,\frac{r_{\text{ctr}}}{\sqrt{d}}\right\rangle \odot \langle +1,-1,\delta_{i_1,i_2}\rangle \odot \cdots \odot \langle +1,-1,\delta_{i_{\zeta-1},i_\zeta}\rangle \odot \left\langle +1,0,\frac{r_{\text{ctr}}}{\sqrt{d}}\right\rangle = \left\langle 0,0,2\frac{r_{\text{ctr}}}{\sqrt{d}} + \sum_{l=1}^{\zeta-1}\delta_{i_l,i_{l+1}}\right\rangle,$$

or

$$\left\langle 0,+1,\frac{r_{\text{ctr}}}{\sqrt{d}}\right\rangle \odot \langle -1,+1,\delta_{i_2,i_1}\rangle \odot \cdots \odot \langle -1,+1,\delta_{i_\zeta,i_{\zeta-1}}\rangle \odot \left\langle -1,0,\frac{r_{\text{ctr}}}{\sqrt{d}}\right\rangle = \left\langle 0,0,2\frac{r_{\text{ctr}}}{\sqrt{d}} + \sum_{l=1}^{\zeta-1}\delta_{i_l,i_{l+1}}\right\rangle.$$

In order to guarantee the feasibility of the second set of loops, we need that $2\frac{r_{\text{ctr}}}{\sqrt{d}} + \sum_{l=1}^{\zeta-1}\delta_{i_l,i_{l+1}} \geq 0$. We derive conditions for the worst case, which occurs when the loop is written for the longest path in T, i.e., when $\zeta = k+1$ and when $\delta_{i_l,i_{l+1}} = -\nu(k)r_{\text{ctr}}$, for all $l \in \{1,\dots,k\}$. In this case, there is no infeasible simple loop if and only if

$$2\frac{r_{\text{ctr}}}{\sqrt{d}} - k\nu(k)r_{\text{ctr}} \geq 0,$$

that is, if and only if $\nu(k) = \frac{2}{k\sqrt{d}}$. □

Finally, the proof of Theorem 3.5 follows from Theorem A.1, Lemma A.2, and Lemma A.3.

Acknowledgment. Giuseppe Notarstefano and Francesco Bullo thank Ruggero Frezza for his kind support.

REFERENCES

[1] G. Notarstefano, K. Savla, F. Bullo, and A. Jadbabaie, *Maintaining limited-range connectivity among second-order agents*, in Proceedings of the IEEE American Control Conference (Minneapolis, MN), 2006, pp. 2124–2129.

[2] H. Ando, Y. Oasa, I. Suzuki, and M. Yamashita, *Distributed memoryless point convergence algorithm for mobile robots with limited visibility*, IEEE Trans. Robotics Automat., 15 (1999), pp. 818–828.

[3] J. Lin, A. S. Morse, and B. D. O. Anderson, *The multi-agent rendezvous problem*, in Proceedings of the 42nd IEEE Conference on Decision and Control (Maui, HI), 2003, pp. 1508–1513.

[4] J. Cortés, S. Martínez, and F. Bullo, *Robust rendezvous for mobile autonomous agents via proximity graphs in arbitrary dimensions*, IEEE Trans. Automat. Control, 51 (2006), pp. 1289–1298.

[5] A. Ganguli, J. Cortés, and F. Bullo, *On rendezvous for visually-guided agents in a nonconvex polygon*, in Proceedings of the 44th IEEE Conference on Decision and Control and European Control Conference (Seville, Spain), 2005, pp. 5686–5691.

[6] D. P. Spanos and R. M. Murray, *Motion planning with wireless network constraints*, in Proceedings of the IEEE American Control Conference (Portland, OR), 2005, pp. 87–92.

[7] M. M. Zavlanos and G. J. Pappas, *Controlling connectivity of dynamic graphs*, in Proceedings of the 44th IEEE Conference on Decision and Control and European Control Conference (Seville, Spain), 2005, pp. 6388–6393.

[8] H. G. Tanner, A. Jadbabaie, and G. J. Pappas, *Flocking in fixed and switching networks*, IEEE Trans. Automat. Control, 52 (2007), pp. 863–868.

[9] Z. Gao, *On discrete time optimal control: A closed-form solution*, in Proceedings of the IEEE American Control Conference (Boston, MA), 2004, pp. 52–58.

[10] B. Aspvall and Y. Shiloach, *A polynomial time algorithm for solving systems of linear inequalities with two variables per inequality*, SIAM J. Comput., 9 (1980), pp. 827–845.

[11] M. Bui, F. Butelle, and C. Lavault, *A distributed algorithm for constructing a minimum diameter spanning tree*, J. Parallel Distrib. Comput., 64 (2004), pp. 571–577.

[12] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Athena Scientific, Belmont, MA, 1997.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 1, pp. 206–233

© 2009 Society for Industrial and Applied Mathematics

CONTROL OF MINIMALLY PERSISTENT FORMATIONS IN THE PLANE*

CHANGBIN YU†, BRIAN D. O. ANDERSON‡, SOURA DASGUPTA§, AND BARIŞ FIDAN‡

Abstract. This paper studies the problem of controlling the shape of a formation of point agents in the plane. A model is considered where the distance between certain agent pairs is maintained by one of the agents making up the pair; if enough appropriately chosen distances are maintained, with the number growing linearly with the number of agents, then the shape of the formation will be maintained. The detailed question examined in the paper is how one may construct decentralized *nonlinear* control laws to be operated at each agent that will restore the shape of the formation in the presence of *small distortions* from the nominal shape. Using the theory of rigid and persistent graphs, the question is answered. As it turns out, a certain submatrix of a matrix known as the rigidity matrix can be proved to have nonzero leading principal minors, which allows the determination of a stabilizing control law.

Key words. multi-agent system, directed formations, distributed control, coordinated motion

AMS subject classifications. 93C10, 68M14

DOI. 10.1137/060678592

1. Introduction and problem description.

The problem of controlling agent formations is gaining more and more attention, as witnessed by an increasing number of contributions in recent years. Among the older contributions, we note, e.g., [1, 2, 3, 4, 5, 6, 7, 8]. Roughly speaking, a collection of agents is prescribed, to move in two- or three-dimensional space, and it is envisaged that they will move as a formation from point A to point B, possibly executing some mission, possibly avoiding obstacles, etc. An agent can be possibly but not necessarily treated as a massless point agent. The words "move as a formation" have the following meaning which a layman might ascribe to them: the formation at one instant of time is congruent to the formation at another instant of time, or equivalently, interagent distances are preserved over all time. Many early contributions deal with the question of just what interagent distances or other constraints are needed to assure this property; see, e.g., [1, 4, 9, 10].

Exactly how motion is achieved in a stable way is an issue of great interest, and recent papers have tended to focus more on the control laws required [11, 12, 13, 14, 15, 16]. It has been observed that if some interagent distances are preserved, for example $2n - 3$ well-chosen distances where n is the number of agents in a two-dimensional formation of point agents, then all interagent distances could be possibly

*Received by the editors December 26, 2006; accepted for publication (in revised form) August 1, 2008; published electronically February 11, 2009. The first author was supported by the Australian Research Council through an Australian Postdoctoral Fellowship under grant DP-0877562. The first, second, and fourth authors were supported by National ICT Australia-NICTA; NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program. The third author was supported by National Science Foundation grants CCF-0729025 and ECS-0622017 and by NICTA.

http://www.siam.org/journals/sicon/48-1/67859.html

†Department of Information Engineering, the Australian National University, Canberra ACT 2600, Australia (Brad.Yu@anu.edu.au).

‡NICTA Ltd. and Department of Information Engineering, the Australian National University, Canberra ACT 2600, Australia (Brian.Anderson@anu.edu.au, Baris.Fidan@anu.edu.au).

§Department of Electrical and Computer Engineering, University of Iowa, Iowa City, IA 52242 (dasgupta@engineering.uiowa.edu).

preserved as a consequence, and a scalable and even distributed control algorithm can be envisaged. Other schemes for control of formation shape can be envisaged too; for example, some angles can be preserved instead of just the distances.

In this paper, like many predecessors, we consider control of formation shape based on interagent distance preservation. What distinguishes this work, however, from most, but not all, work up to this point is that we assign the task of controlling the distance between two agents to a set-point value to only one of the two agents, hence the task is a *directed* one.

Among papers dealing with what one might term *directed* formation control, we note those of [8, 2, 1, 11, 17]. Directed formation control is straightforward if the underlying directed graph is acyclic, since it engenders a triangular coupling, based on a partial ordering of the agents due to an absence of cycles. Thus the challenging problems lie with cyclic graphs. Tabuada, Pappas, and Lima [8] emphasize cyclic graphs, while maintaining a great degree of generality about the nature of the constraints linking the agents. Bailleul and Suri [1] raise the possibility of considering cyclic structures, where there are distance measurements used to achieve control, and argue that such structures are inherently flawed, at least in the presence of noise/bias errors, etc. Lee and Spong [11] consider directed structures, but with the requirement that the underlying graph be balanced (i.e., each node has the same number of inwardly and outwardly directed edges, though a variation is possible with a concept called weighted balancing), and in fact their work is aimed at a different problem (flocking) than preservation of the shape of a two-dimensional formation. Nevertheless, preliminary work of this paper suggests the notion that balanced graphs might also allow efficacious treatment of distance-based formation shape preserving problems, differently to the scheme of this paper.

Earlier on in this work, it was identified that the concept of *graph rigidity* could helpfully underpin much of the control law development. In the undirected graph case, where two agents work together to maintain the correct separation between them, a distributed control law that stabilizes a formation exists only if the underlying graph is rigid, a point specially emphasized in the contributions of Olfati-Saber and colleagues; see, e.g., [5, 2, 3, 4]. In the directed graph case, rigidity is not enough. One needs a further concept, termed *persistence*; see [9, 10]. This concept is reviewed in the next section. It includes rigidity, but overlays this with a further condition that rules out certain information-flow or sensing patterns that are otherwise consistent with the rigidity property. In a persistent graph, it remains possible to have cycles.

The purpose of this work is to demonstrate in detail how control based on distance preservation can be achieved in a directed formation even when cycles are present. We particularly consider a class of such formations termed *minimally persistent formations*; the word *minimal* naturally reflects a certain optimality. *For this class we present a distributed, nonlinear control law and demonstrate its local stability.*

In section 2, we review some background concepts, such as (minimal) rigidity and (minimal) persistence, and introduce different ways of arranging degrees of freedom (DOFs), which eventually affects the control structure. In fact, we focus on formations that have one particular type of DOF distribution. As will be explained later, the specific formations that we address are called "leader-first follower" formations. We then set up the equations of motion for various types of agents, characterized in terms of the number of constraints that each agent maintains. The main results are provided in section 4 and presented by way of verifying a principal minor condition. The structure of eigenvalues is examined in section 5. The paper ends with some concluding remarks in section 6.

2. Background concepts. In this section, we recall a number of graph theoretical concepts related to the maintenance of the shape of a formation. Consider a set of n point agents in the plane. Suppose that they are required to undergo a continuous motion so that the distance between any pair of agents remains constant. As is customary, we will model the formations as graphs: nodes will correspond to the agent positions. An edge will exist between two agents if the distance between these agents is specified as part of the formation specification. We will require that the formation can translate and rotate as a whole, but not flex within itself. This requirement for the formation and its *underlying graph*, viz., the graph modeling it, is called *rigidity*, which is formally defined together with some other fundamental notions of *rigid graph theory* in the following; see, e.g., [18].

Definition 2.1. *In $\Re^2$, a* representation *of an undirected graph $G = (V, E)$ with vertex set V and edge set E is a function $\pi : V \to \Re^2$. We say that $\pi(i) \in \Re^2$ is the* position *of the vertex i and define the distance between two representations π_1 and π_2 of the same graph by $d(\pi_1, \pi_2) = \max_{i \in V} \|\pi_1(i) - \pi_2(i)\|$. A* distance set *$\bar{d}$ for G is a set of distances $d_{ij} > 0$, defined for all edges $(i, j) \in E$. A distance set is* realizable *if there exists a representation π of the graph for which $\|\pi(i) - \pi(j)\| = d_{ij}$ for all $(i, j) \in E$. Such a representation is then called a* realization.

A representation π is rigid *if there exists $\epsilon > 0$ such that for all realizations π' of the distance set induced by π and satisfying $d(\pi, \pi') < \epsilon$, there holds $\|\pi'(i) - \pi'(j)\| = \|\pi(i) - \pi(j)\|$ for all $i, j \in V$ (we say in this case that π and π' are* congruent*). A graph is said to be* generically rigid *(or simply* rigid*) if almost all its representations are rigid.*[1] *A rigid graph is further called* minimally rigid *if no single edge can be removed without losing rigidity.*

Consider a formation F in $\Re^2$ with agents in generic positions and with defined agent pairs having the interagent distances maintained, and let $G = (V, E)$ be the associated graph. Then the formation F is called rigid *if G is rigid. If G is minimally rigid, then F is also called* minimally rigid.

In a rigid formation, it is evident that if enough interagent distances are maintained, then the remainder will be consequentially maintained. For example, see Figure 1; if the distances between the agent pairs (1, 2), (2, 3), (3, 4), (4, 1), and (1, 3) are maintained, then the distance between the pair (2, 4) will be consequentially maintained. For a graph with n agents, it turns out that it is normally enough to maintain $2n-3$ well-distributed distances that are constant in order that all distances are constant. The adverb "normally" connotes that there are some exceptional cases associated with exceptional agent positions. For example, if agents are collinear, or occupy the same position, the conclusion may fail. This conclusion is a standard result of the *rigid graph theory* and is due to [19, 18]. Figure 2 shows an example of two graphs, (a) with well-distributed edges, and (b) without.

Theorem 2.1 (Laman's theorem [19]). *Consider a formation F in $\Re^2$ with agents in generic positions and with defined agent pairs having the interagent distances maintained, and let $G = (V, E)$ be the associated graph. Then F is rigid if and only if there exists a subgraph of G, call it G', with $G' = (V, E')$, $E' \subset E$ such that $|E'| = 2|V| - 3$ and for any $V' \subset V$ defining an induced subgraph $G'' = (V', E'')$ of*

[1] In the graph rigidity literature, the vertex positions in a representation (or the agent positions of a formation) are termed *generic* if the set corresponding to the coordinates of the vertex (or agent) positions is independent over the rationals. An obvious example of nongenericity is when three or more vertices (agents) are collinear. Some discussions on the need for using "generic" and "almost all" can be found in [9, 18].

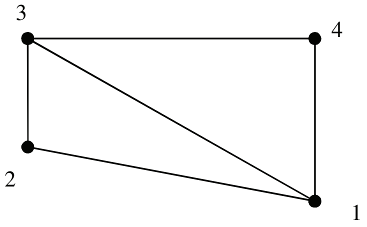


FIG. 1. *A rigid formation in the plane.*

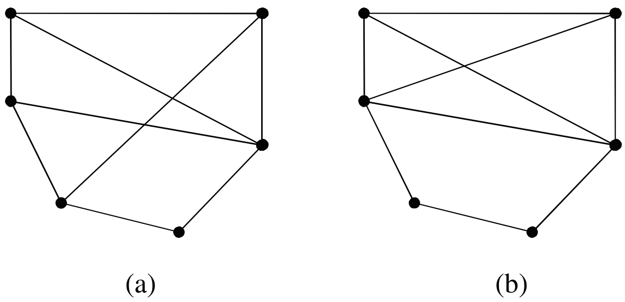


FIG. 2. (a) *Well-distributed edges guarantee rigidity.* (b) *Ill-distributed edges result in nonrigid formation.*

G', there holds $|E''| \leq 2|V'| - 3$.

As a particular immediate corollary of Theorem 2.1, we see that a minimally rigid graph $G = (V, E)$ obeys $|E| = 2|V| - 3$. An alternative characterization of rigidity is provided by the *rigidity matrix*, which we now define.

DEFINITION 2.2. *Order the agents of the formation. Let $p_i \in \Re^2$ be the position of agent i. Each edge has a length $||p_i - p_j||$ for some i, j. Order these edges as $e_1, e_2, \dots,$ then define the edge function $f(p_1, p_2, \dots) = \frac{1}{2}(||e_1||^2, ||e_2||^2, \dots)$. The Jacobian of f is called the* rigidity matrix.

The detailed structure of the rigidity matrix will be important in what follows, and so we record it now. Suppose that the formation has $|V|$ agents and $|E|$ agent pairs with maintained distances (and so there are $|V|$ vertices and $|E|$ edges in the corresponding graph). The rigidity matrix has $|E|$ rows and $2|V|$ columns, with columns $2i - 1$ and $2i$ corresponding to vertex i. When edge i joins vertices j and k, the ith row of the rigidity matrix has four nonzero entries, in columns $2j-1, 2j, 2k-1$, and $2k$ (corresponding to vertices j and k). These entries are, respectively, $x_j - x_k$, $y_j - y_k$, $x_k - x_j$, $y_k - y_j$, where the (x_j, y_j) denotes the coordinates of agent j. The main result is as follows; see [18].

THEOREM 2.2. *Consider a formation F with associated graph $G = (V, E)$, and let the $|E| \times 2|V|$ rigidity matrix be formed as just described. Then for generic*[2] *coordinate values, the formation is rigid if and only if the rigidity matrix has rank $2|V| - 3$.*

Evidently, a minimally rigid formation necessarily has $2|V| - 3$ edges and has a rigidity matrix with full row rank. Evidently also, the kernel of the rigidity matrix has a minimum dimension of 3; any $2n$ vector in the kernel defines both a set of

[2] For a discussion of "generic" in this context, see [18]. Agents would not be in generic positions if, for example, they coincided, or were all collinear.

infinitesimal displacements, which preserve the shape of the formation, and a set of velocity vectors for the agents such that motion along the vector field preserves the interagent distances. For a rigid formation, evidently the kernel has dimension precisely 3. The three independent displacements/motions that it permits correspond to translation in two directions and rotation.

2.1. Constraint consistence and persistence. A distance between two agents can be cooperatively maintained by the two agents, in which case the rigidity ideas can be directly applied. But one can also give the full responsibility for maintaining the constraint to one agent, which has to maintain its distance toward the other agent constant, this latter agent being unaware of that fact and therefore taking no specific action to help satisfy the distance constraint. This unilateral character can be imposed by technical limitations of the autonomous agents or in the interests of greater implementation efficiency. It is this unilateral distance-maintaining approach that is covered in this paper.

Accordingly, henceforth we will deal with directed graphs, where a directed outgoing edge exists from agent i to j if agent i is responsible for maintaining its specified distance from j. The directed edge in this case is denoted by $\{i, j\}$, and the directed graph representing the formation is called the *underlying directed graph* of the formation. As it turns out, a further development of the concept of rigidity is required to understand the idea. Rigidity says that if certain interagent distances are maintained, all other interagent distances are consequentially maintained. We require an additional concept, termed *constraint consistence*, which is equivalent to the requirement that it is possible to maintain the nominated interagent distances. We will require that a formation satisfies both the rigidity and constraint consistence conditions, and if these conditions are satisfied, we will call the formation *persistent.* The notions of *constraint consistence* and *persistence* are formally defined together with some other relevant notions below, following [9].

DEFINITION 2.3. *Consider a directed graph $G = (V, E)$, a set $\bar{d}$ of desired distances $d_{ij} > 0$ for all $\{i, j\} \in E$, and a representation π of G in $\Re^2$. We say that the edge $\{i, j\} \in E$ is* active *if $\|\pi(i) - \pi(j)\| = d_{ij}$. We say that the position of the vertex $i \in V$ is* fitting *for the distance set $\bar{d}$ if it is not possible to increase the set of active edges leaving i by modifying the position of i while keeping the positions of the other vertices unchanged, i.e., if there is no $\pi^* \in \Re^2$ for which $\{\{i, j\} \in E : \|\pi(i) - \pi(j)\| = d_{ij}\} \subset \{\{i, j\} \in E : \|\pi^* - \pi(j)\| = d_{ij}\}$. The realization π is called a* fitting *representation of G for $\bar{d}$ if all the vertices $v \in V$ are at fitting positions for $\bar{d}$. Note that any realization is a fitting representation for its distance set.*

A representation π is called constraint consistent *if there exists $\epsilon > 0$ such that any representation π' fitting for the distance set $\bar{d}$ induced by π and satisfying $d(\pi, \pi') < \epsilon$ is a realization of $\bar{d}$. π is called* persistent *if it is both constraint consistent and rigid (according to Definition* 2.1*, ignoring the edge directions). We say that G is* constraint consistent *(or* persistent*) if almost all its representations are constraint consistent (or persistent, respectively). A persistent graph G is further called* minimally persistent *if it is minimally rigid.*

Consider a formation F in $\Re^2$ with agents in generic positions and with defined agent pairs having the interagent distances maintained and a nominated agent in each pair to maintain the corresponding distance, and let $G = (V, E)$ be the underlying directed graph of F. Then the formation F is called constraint consistent *if G is constraint consistent,* persistent *if G is persistent, and* minimally persistent *if G is minimally persistent.*

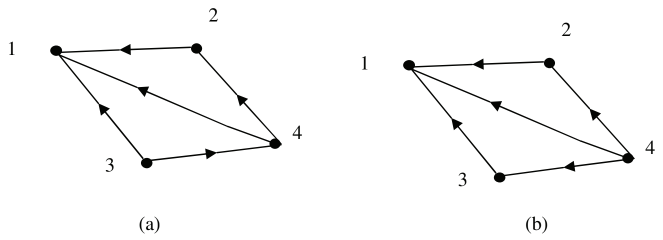


FIG. 3. *Constraint consistent and nonconsistent graphs with the same underlying undirected graph.*

An example of a nonconstraint consistent graph with the underlying undirected graph rigid is shown in Figure 3. If agents 2 and 3 are at their correct distances from agent 1, which allows them some choice of position, agent 4 may not be able to simultaneously achieve its three distance constraints; it cannot of course force agent 2 or 3 to move. In two dimensions, a graph in which the out-degree of every vertex is at most two is automatically constraint consistent, while if the out-degree exceeds two for one or more vertices, it may or may not be constraint consistent [9]. Below we state a particular result regarding minimally persistent graphs (and formations).

THEOREM 2.3 (see [9]). *Consider a directed graph with more than one vertex. Then it is minimally persistent if and only if the underlying undirected graph is minimally rigid and no vertex has more than two outgoing edges.*

It is clear from Theorem 2.3 that in a two-dimensional minimally persistent formation, all agents have zero, one, or two distance constraints to fulfill. Those that have zero distance constraints evidently have two DOFs in which to move, and those with one distance constraint have one DOF. Those with two distance constraints have no DOF. A further requirement for a minimally persistent formation is that the sum of the numbers of DOFs is precisely three, with this in fact corresponding to two translation motions and one rotation motion; then either there is exactly one vertex with two DOFs and another one with one DOF, or there are three vertices, each with one DOF; all other vertices have no DOF.

Nonminimally persistent formations are those where there are extra constraints that have to be fulfilled. Their presence is unnecessary, but may assist in securing robustness against communication or control failures, or they may assist in limiting control magnitudes—just as a linear system with two inputs may be controllable from either input alone but may be much more robustly controlled from both inputs. Not surprisingly, any nonminimally persistent formation contains at least one minimally persistent formation with the same set of vertices and a subset of the edges.

In this paper, we shall confine our attention to the control of minimally persistent formations. Obviously, understanding their control is a precursor to being able to control any formation. Our prime interest is in the following problem. Suppose that all agents in a formation are correctly positioned prior to $t = 0$. Just before $t = 0$, they undergo small displacements from their initial positions. After $t = 0$, those with a DOF are not allowed to exercise that DOF; those required to maintain distances are required to adjust their positions in order to restore any incorrect distances to the correct value. In so doing, they can use only relative position information between themselves and the agents from which they are required to maintain their distance. The whole process has to occur so that the closed loop is stable.

This is a form of zero-input stability. This stability then underpins the preservation of a formation shape when those agents with a positive number of DOFs actually move, so that the whole formation moves while maintaining its shape. In this paper, we do not work through the details of establishing formation shape stability when motion occurs. In this connection, one should recognize that unless the motions executed by the agents with a positive DOF are in some way regular, e.g., constant direction and speed are maintained, one must always expect some distortion of the formation shape away from the nominally correct shape. The analysis in the paper is also a small signal analysis, i.e., we postulate the applicability of linearized models.

2.2. Leader-first follower formations. In this subsection, we categorize minimally persistent formations by the various possible ways DOFs can be allocated. In the rest of the paper we will deal with formations that belong to one such category only, namely, that categorized below as having a leader and first follower structure.

As described in [20], the arrangement of vertices with a positive number of DOFs of a two-dimensional minimally persistent formation can have two possible structures: leader-follower and three-coleaders, and within these structures subtypes occur. The leader-follower structure has two possible subtypes: leader-first follower and leader-remote follower; also the three coleaders in a three-coleader structure may or may not be adjacent to each other, thus giving rise to four possible subtypes: cyclic coleaders, in-line coleaders, one-two coleaders, or distributed coleaders. One can define these terms in more detail as follows.

Definition 2.4. Leader-follower *is a formation structure in which there is an agent (called the* leader*) l that has two DOFs, and another agent f with one DOF. A leader-follower structure is further termed* leader-first follower *if l is a neighbor of f (f is then called the* first follower*) and named* leader-remote follower *if l is not a neighbor of f (f is called the* remote follower *in this case).*

Three-coleaders *is a formation structure in which there are three agents c_1, c_2, c_3 (called the* coleaders*) with one DOF each. A three-coleaders structure is further named* cyclic coleaders *if c_1, c_2, c_3 are adjacent to one another, thus creating a cycle (c_1, c_2, c_3); it is named* in-line coleaders *if the three agents are neighbors but do not form a cycle; it is named* one-two coleaders *if two (say, c_1, c_2) of the three agents are neighbors and the remaining one (c_3) is not a neighbor of c_1 or c_2; and it is named distributed coleaders, if none of c_1, c_2, c_3 are neighbors of one another.*

In this paper, the formations we study *are of a leader-first follower variety.* Figure 4 provides an example of such a formation. Note in particular that 5 is the first follower and 6 is the leader. All the other agents have exactly two outgoing edges (thus guaranteeing minimal persistence, as discussed in the previous subsection), and there is exactly one cycle {1,2,3}. There exists a formation with this same graph, but the coordinates are chosen such that an intuitively appealing control law presented later actually will lead to instability, as shown by an example later in section 5.

3. Equations of motion. In this section, we set up the equations related to the control of a minimally persistent formation in the plane. The formation has a leader and first follower. To motivate the derivation of the equations, we will first describe the approach using discrete time ideas. The derivation itself will, however, use continuous time.

With a leader-first follower structure, there are two vertices with a positive number of DOFs, and the remainder have no DOFs, having to maintain their distance from precisely two other vertices.

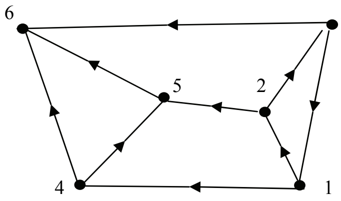


FIG. 4. *An example of a minimally persistent formation with leader-first follower structure and a cycle* $\{1, 2, 3\}$.

A discrete time view of the adjustment process is as follows. Prior to time 0, all distances are correct. At time 0, all agents are moved a small amount, so that distances are no longer correct. Between time 0 and 1 all agents determine the position of the agents from which they are required to maintain a correct distance, and then determine the point to which they would have to move to correct any distance errors. The leader itself makes no determination and remains permanently stationary after its initial move at time 0. Between time 0 and 1, the first follower determines the point on a straight line joining it to the leader which is at the correct distance from the leader. There are two possible points to which all other agents could move; each chooses the point closest to their position just after time 0.

At time 1, all agents then move to the positions they determined between 0 and 1. Between time 1 and 2, all agents review whether their positions are correct by checking the current distances to their neighbors, and determine the correction that would be required to re-establish the required distances, assuming their neighbors do not move. At time 2 the mispositioned agents actually execute a move. This process can clearly be repeated.

In the case of formations with an acyclic graph, it is clear that after a finite number of steps, all agents will become stationary. Indeed, after r steps, those agents with a path of length at most r steps to the first follower and the leader will cease to move. In the case of a formation with a cyclic graph, the question arises as to whether the process of adjustment is of infinite duration (it is, in general), and if so, whether the adjustments get smaller and smaller so that the agent positions converge, or whether there is continuous oscillation in the position. Our focus is on this case.

3.1. General approaches to control law derivation for formations. As is common, though not universal [1, 12, 13, 21, 22, 23], we shall adopt a simple kinematic velocity control model for each agent:

$$\dot{p}_i = u_i. \tag{3.1}$$

Suppose that agent i is tasked with maintaining distances d_{ij}^* and d_{ik}^* from two other agents j and k. The closed-loop control laws will typically be of the form

$$u_i = u_i(p_j - p_i, p_k - p_i, d_{ij}^*, d_{ik}^*). \tag{3.2}$$

There are several key points we need to make about this law. First, the law uses *relative positions* of agents j and k, and not just the current distances of agent i from agents j and k. Thus *more needs to be sensed than is controlled*, a not uncommon situation in modern control. Relative positions can be sensed if agent i is equipped with distance and direction sensors; alternatively, if agents are able to sense distances not just of their neighbors in the graph we are using to define the formation, but of their two-hop neighbors in the same graph, and if they are able to pass those distances to their neighbors, relative positions can be sensed. Thus if agent j can sense its distance from agent k and inform agent i, then agent i can determine the angle between the lines joining it to agents j and k. If agent i has its own coordinate basis, then knowledge of this angle is equivalent to knowledge of relative positions.

Agent i needs also to know not just the values of d^*_{ij} and d^*_{ik} but also the orientation of the triangle $i - j - k$ when agents are in their correct positions, and it is assumed that the departures from equilibrium conditions are sufficiently small that the orientation of this triangle is not disturbed.

Second, the law has to have a rotational invariance property. If R_i is a rotation matrix, it is clear that we need the property

$$u_i(R_i(p_j - p_i), R_i(p_k - p_i), d^*_{ij}, d^*_{ik}) = R_i u_i(p_j - p_i, p_k - p_i, d^*_{ij}, d^*_{ik}).$$

This expresses the fact that if the coordinate system in which positions are measured is rotated, the same rotation needs to apply to the controls, which determine derivatives of position. A consequence of this fact is that each agent can compute using its own local coordinate system, and agents do not need to share a common understanding of the direction of north. To see this, let q_{ij} denote the position of agent j in some local coordinate system maintained by agent i. Then there exists a rotation matrix R_i and a translation vector τ_i such that

$$q_{ij} = R_i p_j + \tau_i.$$

If the control is computed using the local coordinate basis, it will be

$$\begin{aligned} u_i(q_{ij} - q_{ii}, q_{ik} - q_{ii}, d^*_{ij}, d^*_{ik}) &= u_i(R_i(p_j - p_i), R_i(p_k - p_i), d^*_{ij}, d^*_{ik}) \\ &= R_i u_i(p_j - p_i, p_k - p_i, d^*_{ij}, d^*_{ik}) \end{aligned} \tag{3.3}$$

and in the global coordinate basis, this is the same as (3.2).

Third, the control law is decentralized. Obviously, its implementation just uses sensed data local to agent i. It could also have been decentralized in a second sense, specifically if the design of the law for agent i apparently took no account of the design for other agents. As it turns out, to control a persistent formation, we are able to propose a law which is decentralized in its operation, but *not* decentralized in its design. Put another way, the control law we end up proposing for agent i will depend on more data defining the desired formation shape (but not the current formation shape) than just the two distances d^*_{ij} and d^*_{ik}, even though the only data from the current formation shape are the relative positions of its neighbors. We will in fact present an approach which fixes the control laws for each agent in sequence, and the law for any one agent depends on the parameters of the laws for the preceding agents of the sequence. The sensed data, however, for the law at each agent is unchanged.

As for the determination of the actual law, it is common [1, 2, 3, 4, 5, 6, 7, 12, 21] when distance constraints are bidirectional to select a type of Lyapunov function, typically reflecting the distance errors, and to choose the control laws to ensure that the derivative is nonpositive. (It is generally not possible to ensure that the derivative is negative). Given that the Lyapunov function converges to a limit, it is a separate issue to show that this convergence implies correct convergence of the formation shape. This approach, however, does not really extend easily to the case where the distance

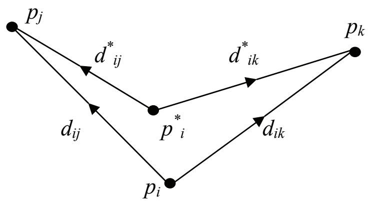


FIG. 5. *Illustration of position adjustment for point agent i at p_i with respect to agent j at p_j and agent k at p_k. The desired position is at p_i^*.*

constraints are unidirectional, which is the reason motivating the alternative approach outlined at the start of the section.

Last, we note that it is common to distinguish between results applicable to linearized closed-loop models, and results which offer convergence starting from a very wide range of initial conditions. Very few "almost global" results are actually available [21, 22, 23], serving as examples of exceptions. In this paper, we set up a nonlinear law and prove convergence of a linearized version in which adjustable parameters have to be set at certain values. Of course, if for a nonlinear law, convergence of a linearized version is proved, it may be that convergence for the nonlinear law will occur for a wide range of initial conditions, but in general this has to be established with an investigation of a particular case.

3.2. A nonlinear law. Suppose the agents are numbered from 1 to n, with the first follower and leader $n-1$ and n, respectively. Suppose that the initial (prior to time 0) position of all vertices is given (in a global coordinate basis) by $p_{i0} = [x_{i0}\ y_{i0}]'$, $i = 1, 2, \dots, n$, with distance constraints all satisfied. Suppose that at time 0, all agents are displaced from their initial positions. In a moment, we shall impose a bound on the magnitude of the displacement.

At time t, agent i uses the relative position information of its neighbors (suppose they are agents j and k) to determine the point $p_i^*(t)$ which is at the correct distances d_{ij}^* and d_{jk}^* from agents j and k, respectively, and (noting that there are two such points) is the closer of the two possible points to p_i. (See Figure 5.) In order to do this, it is required that the displacements of p_j and p_k from their initial positions not be so great as to mean there is no possible point p_i^*. If the agents p_j and p_k were to remain at their initial predisplacement positions, p_i^* would coincide with the initial predisplacement position p_i of agent i. If they were to move so that their separation exceeded $d_{ij}^* + d_{jk}^*$, no p_i^* could be found. Obviously then, a nonzero upper bound on the displacements can be found, assuring the existence of p_i^*. Observe that we can write

$$p_i^* - p_i = f(p_j - p_i, p_k - p_i, d_{ij}^*, d_{ik}^*) \tag{3.4}$$

for some function f which is independent of i. The control law to be used is one which moves p_i closer to p_i^*, but it makes no allowance for the fact that because p_j and p_k are likely to be moving, p_i^* will also be changing. Thus we suppose that for some K_i,

$$\dot{p}_i = K_i(p_i^* - p_i) = K_i f(p_j - p_i, p_k - p_i, d_{ij}^*, d_{ik}^*). \tag{3.5}$$

If p_i^* were to be constant, any K_i with positive real part eigenvalues could be used, including $K_i = I$. However, p_i^* will not be constant in general, and so a more sophisticated way of choosing K_i is needed. It will be shown in section 5 by way of an example that the choice $K_i = I$ for all i may actually be destabilizing, and indeed the example displays stabilizing gains for which at least one of the K_i's does not even have positive real part eigenvalues. A significant part of the paper from subsection 3.5 will deal with the basis for choosing K_i. As indicated in the previous subsection, the law in question will be decentralized in operation, i.e., only local sensed data are used, but in design, it will not be decentralized; for, as it turns out, the choice of the 2×2 gain matrices K_i to assure stability demands this.

Equation (3.5) covers agents 1 through $n-2$. Agent n, the leader, will be assumed stationary. The law for agent $n-1$, the first follower, is one which requires it to determine p^*_{n-1} as the point on the line joining p_{n-1} to p_n which is at the correct distance for the first follower from p_n, and then the first follower moves toward that point.

Thus we have

$$p^*_{n-1} - p_{n-1} = \frac{||p_n - p_{n-1}|| - d^*_{n-1,n}}{||p_n - p_{n-1}||}(p_n - p_{n-1}), \tag{3.6}$$

and for the control law, with some positive k_{n-1},

$$\dot{p}_{n-1} = k_{n-1}(p^*_{n-1} - p_{n-1}) = k_{n-1}\frac{||p_n - p_{n-1}|| - d^*_{n-1,n}}{||p_n - p_{n-1}||}(p_n - p_{n-1}). \tag{3.7}$$

Together with

$$\dot{p}_n = 0, \tag{3.8}$$

(3.5) and (3.7) define the nonlinear closed-loop system. Rather than giving a formal proof of existence of solutions, later in the paper we shall demonstrate that stability can be assured for the linearized equations through an appropriate choice of the K_i and k_{n-1}, which is an indirect proof of solution existence.

3.3. Linearized equations. Let us suppose henceforth that all displacements are small enough to allow first order approximation, and in particular that we can represent at all times the position of agent i by $p_i(t) = \delta p_i(t) + \bar{p}_i$, where the $\bar{p}_i$ correspond to agent positions for which all desired distance constraints are met, and $\delta p_i(t)$ is small. Let $p_i(t) = [x_i(t)\ y_i(t)]'$, $\bar{p}_i = [\bar{x}_i\ \bar{y}_i]'$, and $\delta p_i(t) = [\delta x_i(t)\ \delta y_i(t)]'$. Below, we will indicate more specifically how the $\bar{p}_i$ are determined. Note that $p^*_i(t) \neq \bar{p}_i$ in general. This is because p^*_i would denote an equilibrium position for p_i only if p_j and p_k never moved. In general they will move. We assume also that all quantities $||p_i(t) - p^*_i(t)||$ are small, which can be guaranteed if the initial displacements away from equilibrium are all small and the subsequent motion is stable.

We consider first agents 1 through $n-2$. Refer to Figure 5, and apply the cosine law to the triangle with corners p_i, p^*_i, and p_j. Because $||p_i - p^*_i||$ is small, there holds (neglecting the square of $||p_i - p^*_i||$)

$$||p_j - p_i||^2 - 2[p_j - p_i]^T[p^*_i - p_i] \approx ||p_j - p^*_i||^2, \tag{3.9}$$

which may be rewritten as

$$2[p_j - p_i]^T[p^*_i - p_i] \approx d^2_{ij} - d^{*2}_{ij}.$$

Noting that $p_j - p_i = \bar{p}_j - \bar{p}_i + \delta p_j - \delta p_i$, and again neglecting second order terms, we get the further approximation

$$2[\bar{p}_j - \bar{p}_i]^T[p_i^* - p_i] \approx d_{ij}^2 - d_{ij}^{*2}. \tag{3.10}$$

Accordingly, provided agents i, j, and k are not collinear (i.e., that the vector $p_i - p_j$ is not parallel to the vector $p_i - p_k$), we will have

$$p_i^* - p_i \approx \frac{1}{2}\begin{bmatrix} \bar{x}_j - \bar{x}_i & \bar{y}_j - \bar{y}_i \\ \bar{x}_k - \bar{x}_i & \bar{y}_k - \bar{y}_i \end{bmatrix}^{-1}\begin{bmatrix} d_{ij}^2 - d_{ij}^{*2} \\ d_{ik}^2 - d_{ik}^{*2} \end{bmatrix}. \tag{3.11}$$

It is normal in applying rigid graph theory to formations to assume that the formations are generic; a consequence of this assumption is that colinearities are excluded, and so the matrix inverse in (3.11) exists.

Next, it is straightforward to check that, again neglecting second order terms,

$$\begin{aligned} \frac{1}{2}[d_{ij}^2 - d_{ij}^{*2}] &= \frac{1}{2}[||p_j - p_i||^2 - ||\bar{p}_i - \bar{p}_j||^2] \\ &= \frac{1}{2}[||(\bar{p}_j - \bar{p}_i) + (\delta p_j - \delta p_i)||^2 - ||\bar{p}_j - \bar{p}_i||^2] \\ &\approx [\bar{p}_i - \bar{p}_j]^T\delta p_i - [\bar{p}_i - \bar{p}_j]^T\delta p_j, \\ \frac{1}{2}[d_{ik}^2 - d_{ik}^{*2}] &\approx [\bar{p}_i - \bar{p}_k]^T\delta p_i - [\bar{p}_i - \bar{p}_k]^T\delta p_k. \end{aligned}$$

Putting this together with (3.5), there results

$$\begin{bmatrix} \delta\dot{x}_i \\ \delta\dot{y}_i \end{bmatrix} = K_i \begin{bmatrix} \bar{x}_j - \bar{x}_i & \bar{y}_j - \bar{y}_i \\ \bar{x}_k - \bar{x}_i & \bar{y}_k - \bar{y}_i \end{bmatrix}^{-1} R_{(ij,ik)} \begin{bmatrix} \delta x_i \\ \delta y_i \\ \delta x_j \\ \delta y_j \\ \delta x_k \\ \delta y_k \end{bmatrix} \tag{3.12}$$

with $R_{(ij,ik)} = \begin{bmatrix} \bar{x}_i - \bar{x}_j & \bar{y}_i - \bar{y}_j & -\bar{x}_i + \bar{x}_j & \bar{y}_j - \bar{y}_i & 0 & 0 \\ \bar{x}_i - \bar{x}_k & \bar{y}_i - \bar{y}_k & 0 & 0 & -\bar{x}_i + \bar{x}_k & -\bar{y}_i + \bar{y}_k \end{bmatrix}.$

Observe for later reference that $R_{(ij,ik)}$ is a submatrix of the rigidity matrix, with rows corresponding to the edges $\{i, j\}$ and $\{i, k\}$ and columns corresponding to vertices i, j, and k.

We turn now to the equation governing the first follower. Consider first the motion defined by (3.7), and recall that the first follower moves along the line joining it to the leader. Further, the leader remains stationary, and it is logical then to take $\bar{p}_n$ as its position so that $p_n(t) = \bar{p}_n$. Then the instantaneous target point for the first follower remains constant, and it is natural to take this point as the desired equilibrium position, i.e., $p_{n-1}^*(t) = \bar{p}_{n-1}$.

For the linearized system, there results

$$\begin{bmatrix} \delta\dot{x}_{n-1} \\ \delta\dot{y}_{n-1} \end{bmatrix} = k_{n-1}I_2 \begin{bmatrix} \bar{x}_n - \bar{x}_{n-1} & \bar{y}_n - \bar{y}_{n-1} \\ -(\bar{y}_n - \bar{y}_{n-1}) & \bar{x}_n - \bar{x}_{n-1} \end{bmatrix}^{-1} R_{((n-1)n,00)} \begin{bmatrix} \delta x_{n-1} \\ \delta y_{n-1} \\ \delta x_n \\ \delta y_n \end{bmatrix} \tag{3.13}$$

with

$$R_{((n-1)n,00)} = \begin{bmatrix} \bar{x}_{n-1} - \bar{x}_n & \bar{y}_{n-1} - \bar{y}_n & -\bar{x}_{n-1} + \bar{x}_n & -\bar{y}_{n-1} + \bar{y}_n \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Of course, the equations for the leader are

$$\begin{bmatrix} \dot{\delta x}_n \\ \dot{\delta y}_n \end{bmatrix} = 0. \tag{3.14}$$

In the light of the above discussion, one could arrive at the following theorem which is the main result of this section.

THEOREM 3.1. *The linearization of* (3.5), (3.7), *and* (3.8) *under the first order approximation is*

$$\dot{\delta p}(t) = K R_e^{-1} \begin{bmatrix} R \\ 0 \end{bmatrix} \delta p(t), \tag{3.15}$$

where K and R_e are diagonal block matrices, with each block of size 2×2. (The last block of R_e for convenience can be taken as the identity).

Of course, $K = \mathrm{diag}[K_1, K_2, \dots, K_{n-2}, k_{n-1} I_2, 0]$, and the first $(n-2)$ diagonal blocks of R_e are obtained as 2×2 submatrices of the rigidity matrix, selecting rows corresponding to edges $\{i, j\}, \{i, k\}$ and columns corresponding to vertex i.

3.4. Simplified dynamics. Our ultimate goal is to show that through a suitable choice of gains, the nonlinear system can be stabilized. This will be done by choosing gains to stabilize the system obtained by linearizing the nonlinear system around the equilibrium point. However, as is evident from the linearized equation (3.15), there will necessarily be three modes of the linearized system which are located at the origin; this apparently makes it much more difficult to establish a stability result for the nonlinear system than for the linear system. Nevertheless, a modest modification of the usual approach will work.

For the purpose of the theoretical analysis, without loss of generality let us choose the global coordinate basis so that the x-axis coincides with the line joining agents $n-1$ and n at the start of the motion.[3] Because agent $n-1$ moves solely on this line, it will stay on the x-axis, and so there will hold $y_{n-1}(t) = \bar{y}_{n-1} = \bar{y}_n$ for all t. Obviously then, for the nonlinear system, $\delta y_{n-1}(t) = y_{n-1}(t) - \bar{y}_{n-1}$ will be identically zero. Because the leader does not move, for the nonlinear system δx_n and δy_n are identically zero.

Examination of the linearized equation (3.13) shows also that

$$\begin{bmatrix} \dot{\delta x}_{n-1} \\ \dot{\delta y}_{n-1} \end{bmatrix} = \begin{bmatrix} -k_{n-1} \delta x_{n-1} \\ 0 \end{bmatrix}, \tag{3.16}$$

while the equations for the leader remain the same as (3.14). It is straightforward to check that the second entry of (3.16) is also true for the original nonlinear system.

Let $\hat{R}_e$ denote R_e with the last three rows and columns discarded; i.e., it is a $(2n-3) \times (2n-3)$ submatrix of R_e. Correspondingly, we have to use a $(2n-3) \times (2n-3)$ submatrix of K and consider $K = \mathrm{diag}[\hat{K}, 0_3]$, i.e.,

$$\hat{K} = \left(\bigoplus_{i=1}^{n-2} K_i \right) \bigoplus k_{n-1}, \tag{3.17}$$

[3] If the new global basis is obtained from the original one by a rotation and translation, the gains K will all differ in the two coordinate bases by the same orthogonal similarity transformation.

where $\hat{K}_i$ are each 2×2 and k_{n-1} is a scalar.

Recall also that

$$\begin{bmatrix} \delta y_{n-1} \\ \delta x_n \\ \delta y_n \end{bmatrix} = 0. \tag{3.18}$$

This means that we can replace (3.15) with the simplified equation

$$\begin{bmatrix} \dot{\delta x_1} \\ \dot{\delta y_1} \\ \dot{\delta x_2} \\ \dot{\delta y_2} \\ \vdots \\ \dot{\delta x_{n-2}} \\ \dot{\delta y_{n-2}} \\ \dot{\delta x_{n-1}} \end{bmatrix} = \hat{K}\hat{R_e}^{-1}\hat{R} \begin{bmatrix} \delta x_1 \\ \delta y_1 \\ \delta x_2 \\ \delta y_2 \\ \vdots \\ \delta x_{n-2} \\ \delta y_{n-2} \\ \delta x_{n-1} \end{bmatrix}, \tag{3.19}$$

where $\hat{R}_e$ and $\hat{R}$ are obtained by removing the last three columns from R_e and R, respectively.

Notice that (3.18) also holds for the nonlinear system. Equation (3.19) is actually the linearized version of the nonlinear equations governing the first $2n-3$ coordinates. Accordingly, if $\hat{K}$ in (3.19) is chosen to ensure that (3.19) is exponentially stable, then the nonlinear equations governing the first $2n-3$ coordinates will be exponentially stable for all initial conditions within some domain of attraction, and the motion of the last three coordinates is trivially defined by (3.18), i.e., there is no motion.

Observe in (3.19) that $\dot{\delta p_i}$ can depend on only $\delta p_i, \delta p_j$, and δp_k, where agents j and k are those from which agent i must maintain its distance. This forces $\hat{K}$ to have the structure of (3.17), but does not constrain the individual blocks to be, for example, individually diagonal or multiples of the identity. For this problem, $\hat{K}$ in fact serves as the controller. Of course, one could contemplate replacing $\hat{K}$ by some dynamics, in which $\dot{\delta p_i}$ was determined by dynamic processing of $\delta p_i, \delta p_j$, and δp_k, but this is beyond the scope of the paper.

It should be noted that in (3.19), the matrix $\hat{R}_e^{-1}\hat{R}$ is defined in terms of the target positions $\bar{p}_i$, which of course are not known a priori. Consider now, though, the setting where a previously intact formation was deformed by the "small" movements of multiple agents. *Then if under a given $\hat{K}$, $\hat{K}\hat{R}_e^{-1}\hat{R}$ is sufficiently Hurwitz, with $\bar{p}_i$ in $\hat{R}_e^{-1}\hat{R}$ replaced by the positions prior to the deformation, then under sufficiently small movements that cause the deformation,* (3.19) *will remain Hurwitz. Thus, in what follows, which is concerned with designing $\hat{K}$ to ensure local stability, to avoid notational complexities we will assume that the matrix $\hat{R}_e^{-1}\hat{R}$ is formed with the positions prior to deformation replacing $\bar{p}_i$, subject to the coordinate transformation described in this section.*

3.5. Choosing the block diagonal control multiplier. From the above analysis of a continuous time version of the formation shape maintenance problem, we have determined that the underlying dynamic equation is of the form

$$\dot{z} = \Lambda A z. \tag{3.20}$$

In this equation, Λ is a diagonal or possibly block diagonal matrix. Its entries correspond to gains associated with the control used by each agent. If each agent were to apply the same gain to the two distance constraints, then we would have Λ of the form $\lambda_1 I_2 \oplus \lambda_2 I_2 \oplus \dots$.

For the moment, in this section we will assume that Λ is diagonal and all the diagonal elements of Λ can be independently chosen. The key result, which will be established with a constructive procedure, is as follows.

THEOREM 3.2. *Suppose A is an $m \times m$ nonsingular matrix with every leading principal minor nonzero. Then there exists a diagonal Λ such that the real parts of the eigenvalues of ΛA are all negative.*

Proof. The proof of the theorem will proceed by induction on m. We shall ensure that $m - 1$ eigenvalues of ΛA lie very far in the left half plane by selecting the first $m - 1$ diagonal entries of Λ, and then show how the mth diagonal entry λ_m can be chosen to make the last eigenvalue of ΛA have negative real part. Suppose the theorem is true for $m = 1, 2, \dots, r - 1$ (it is trivially true for $m = 1$). Consider the case $m = r$ and suppose A has nonzero leading principal minors. Write

$$A = \begin{bmatrix} A_{11} & a_{12} \\ a_{21}^T & a_{22} \end{bmatrix},$$

where A_{11} is $(r - 1) \times (r - 1)$ and nonsingular with nonsingular leading principal minors, $a_{12}, a_{21} \in \Re^{r-1}$, $a_{22} \in \Re$. Appealing to the induction hypothesis, choose Λ_1 diagonal so that $\Lambda_1 A_{11}$ has all eigenvalues with negative real parts. Now recall that if

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \end{bmatrix} = \begin{bmatrix} \epsilon^{-1}\Lambda_1 A_{11} z_1 + \epsilon^{-1}\Lambda_1 a_{12} z_2 \\ \lambda_2 a_{21}^T z_1 + \lambda_2 a_{22} z_2 \end{bmatrix}$$

and if the real parts of the eigenvalues of $\Lambda_1 A_{11}$ are all negative, then provided ϵ is small enough we can use singular perturbation theory to study stability [24]. The high order system is asymptotically stable if an associated lower order system is asymptotically stable. This low order system is obtained by replacing the differential equation for z_1 by the equation

$$\begin{aligned} z_1 &= -(\epsilon^{-1}\Lambda_1 A_{11})^{-1}(\epsilon^{-1}\Lambda_1 a_{12}) z_2 \\ &= -A_{11}^{-1} a_{12} z_2, \end{aligned}$$

and then the differential equation for z_2 becomes

$$\begin{aligned} \dot{z}_2 &= -\lambda_2 a_{21}^T A_{11}^{-1} a_{12} z_2 + \lambda_2 a_{22} z_2 \\ &= \lambda_2 [a_{22} - a_{21}^T A_{11}^{-1} a_{12}] z_2. \end{aligned}$$

Choosing λ_2 so that $\lambda_2[a_{22} - a_{21}^T A_{11}^{-1} a_{12}] < 0$ ensures stability of z_2 and then the whole of z. Also,

$$|A| = |A_{11}|(a_{22} - a_{21}^T A_{11}^{-1} a_{12}),$$

so there is no possibility that $a_{22} - a_{21}^T A_{11}^{-1} a_{12} = 0$. □

Remark 3.1. Examining the literature for theorems of this variety, one can see that if A and Λ are permitted to be complex, then the same leading principal minor condition guarantees that a Λ can be found to produce any prescribed set of eigenvalues; see [25]. However, the method of proof cannot be carried over to the real case. Further, it is easy to show with a 2×2 counterexample that eigenvalue positionability in the real case cannot always be guaranteed, even with nonzero principal minors.

Remark 3.2. In order to apply this theorem to our situation, we will need to show that the matrix $\hat{R}$ defined in the previous subsection (which corresponds to A in the theorem) has all principal minors nonzero. This is not straightforward and will be done in the next section. Of course, $\hat{K}(\hat{R}_e)^{-1}$ will correspond to Λ, and we achieve a Λ by choice of $\hat{K}$ that in turn conforms to (3.17).

Remark 3.3. The condition of the theorem is a sufficiency condition. It would appear not to be necessary. However, in order that a stabilizing Λ exist, it is certainly necessary that A be nonsingular. For if A is singular, then so is ΛA and there is necessarily at least one zero eigenvalue. Also, suppose A is $m \times m$, and the characteristic polynomial is $|sI - A| = A^m + \alpha_1 A^{m-1} + \cdots + \alpha_m$; then $\alpha_i = (-1)^i \sum$ (all $i \times i$ principal minors of A). Hence if all $i \times i$ principal minors of A are zero, then $\alpha_i = 0$. Also, all $i \times i$ principal minors of A are zero if and only if all $i \times i$ principal minors of ΛA (for nonsingular Λ) are zero. Hence if all $i \times i$ principal minors of A are zero, the characteristic polynomial of ΛA will have a zero coefficient. Since for the characteristic polynomial to be stable, all coefficients must be positive, we see that a necessary condition for Λ to exist is that for all i, at least one $i \times i$ principal minor of A is nonzero.

Remark 3.4. There is a trivial extension to the theorem: if the rows and columns of A can be symmetrically reordered so that every leading principal minor is nonzero, then Λ can be chosen with the desired properties.

4. The principal minor condition.

Recall from the previous sections that the key technical condition required for stabilizability is that a certain matrix have all its leading principal minors nonzero. This section addresses this issue. In particular, recall that with $V = \{1, \ldots, n\}$, the directed graph $G = (V, E)$ has a leader-first follower structure with n and $n-1$ being the leader and the first follower, respectively. In what follows, suppose the rigidity matrix R is such that its last row corresponds to the only outgoing edge that the first follower has, i.e., the outgoing edge from $n-1$ to n.

Recall that $\hat{R}$ is the $(2n-3) \times (2n-3)$ submatrix of the rigidity matrix R of G, obtained by removing the last three columns of R. Further, $\hat{R}_e$ is the $(2n-3)\times(2n-3)$ matrix as below:

$$\left(\bigoplus_{i=1}^{n-2} B_i\right) \bigoplus (x_n - x_{n-1}), \tag{4.1}$$

where if node $i \in V' = \{1, \ldots, n-2\}$ has outgoing edges to j and k, then

$$B_i = -\begin{bmatrix} x_i - x_j & y_i - y_j \\ x_i - x_k & y_i - y_k \end{bmatrix}. \tag{4.2}$$

This implicitly enforces the following ordering of the rows and columns of $\hat{R}$. Columns $2k-1$ and $2k$ correspond to node $k \in V'$. Rows $2k-1$ and $2k$ correspond to the two outgoing edges of node $k \in V'$. The last row of R corresponds to the edge from $n-1$ to n.

Our goal is to prove the following main result of this section.

THEOREM 4.1. *Consider an n-node minimally persistent formation F with agent set $P = \{1, \ldots, n\}$ at generic positions, and n and $n-1$ the leader and the first follower, respectively. Suppose $\hat{R}$ is the $(2n-3) \times (2n-3)$ submatrix of the rigidity matrix R of F, obtained by removing the last three columns of R and obeying the row and column ordering noted above. Then there exists an ordering of the first $n-2$*

vertices of F and an ordering of the pair of outgoing edges for each of these vertices such that the leading principal minors of the associated $\hat{R}$ are generically nonzero.

In order to prove this theorem, we will first establish several lemmas and another theorem in this section. The proof is completed at the end of the section. The significance of this result is as follows. From Theorem 3.2, one can find a diagonal Λ such that $\Lambda\hat{R}$ has all eigenvalues in the open left half plane. Consequently, with

$$\hat{K} = \Lambda\hat{R}_e \tag{4.3}$$

the matrix $\hat{K}\hat{R}_e^{-1}\hat{R}$ can be made to have all eigenvalues in the open left half plane. For such a choice the control laws of the previous section stabilize the system. Further, the the block diagonal nature of $\hat{R}_e$ outlined in (4.1) and (4.2) ensures that $\hat{K}$ has the right structure, in that it is block diagonal with the first $n-2$ diagonal blocks being 2×2 matrices and the last diagonal element being scalar. As noted in the previous section, this ensures that to implement its individual control law each agent need only sense the relative positions of its neighbor agents.

To proceed with the proof of Theorem 4.1 we first prove that the matrix $\hat{R}$, obtained by removing the last three columns of the rigidity matrix R, is generically nonsingular.

Lemma 4.2. *Under the hypothesis of Theorem* 4.1, *$\hat{R}$ is generically nonsingular.*

Proof. The null space of R has as its basis the following three $2n$-vectors:
$\eta_1 = [1, 0, 1, 0, 1, 0, \ldots]'$,
$\eta_2 = [0, 1, 0, 1, 0, 1, 0, \ldots]'$, and
$\eta_3 = [y_1, -x_1, y_2, -x_2, y_3, -x_3, \ldots, y_{n-1}, -x_{n-1}, y_n, -x_n]'$.
If the $(2n-3) \times (2n-3)$ matrix $\hat{R}$ is singular, it must have a $(2n-3)$-dimensional null vector $\eta \neq 0$. Then

$$R[\eta', 0, 0, 0]' = 0.$$

Thus as $[\eta', 0, 0, 0]'$ is in the space spanned by η_i, and $\eta \neq 0$, the matrix formed by the last three elements of each of η_i,

$$\begin{bmatrix} 1 & 0 & -x_{n-1} \\ 0 & 1 & y_n \\ 1 & 0 & -x_n \end{bmatrix},$$

must be generically singular. This is of course false. □

We require further notation. Recall that the set V' comprises the follower nodes. Consider now a subset of the follower nodes: $V_1 \subset V'$. Then we define $R(V_1)$ as the principal submatrix of $\hat{R}$ obtained by retaining the columns corresponding to the elements of V_1. Call $G_1 = (V_1, E_1)$ the subgraph of G induced by V_1 and conforming to the row and column ordering noted earlier. Note $R(V_1)$ is not the rigidity matrix of the induced subgraph G_1, as it may contain edge information regarding certain edges of G which are not in G_1.

First we have the following lemma.

Lemma 4.3. *Under the hypothesis of Theorem* 4.1, *$R(V')$ is generically nonsingular.*

Proof. This follows by noting that with $\times$ a *don't care* vector, $\hat{R}$ can be partitioned as

$$\hat{R} = \begin{bmatrix} R(V') & \times \\ 0 & x_{n-1} - x_n \end{bmatrix}.$$

Then the result follows from Lemma 4.2. □

Next, we prove the following lemma.

Lemma 4.4. *Consider a minimally persistent graph with a leader-first follower structure $G = (V, E)$, and any induced subgraph $G_1 = (V_1, E_1)$, such that V_1 contains neither the leader nor the first follower. Call $V_2 \subset V - V_1$ the set of vertices in $V - V_1$ that have incoming edges from V_1 in G. Suppose $|V_2| \geq 2$. Define E_{21} as the set of outgoing edges of nodes of V_1 in G and terminating in V_2. Construct a new graph $\bar{G} = (\bar{V}, \bar{E})$ with the following properties:*

(a) $\bar{V} = V_1 \cup V_2$.

(b) *Let $\widetilde{G}_2 = (V_2, \widetilde{E}_2)$ be any minimally persistent graph with a leader-first follower structure, with vertex set V_2, and with edge set $\widetilde{E}_2$ that is not required to be related in any way to the edges in G. Choose $\bar{E} = \widetilde{E}_2 \cup E_{21} \cup E_1$.*

Then $\bar{G}$ is a minimally persistent graph with a leader-first follower structure, and the leader and first follower belong to V_2.

Proof. By construction, while all outgoing edges of nodes in V_1 in the original graph G are in $\bar{E}$, no incoming edge to V_1 from V_2 in E also appears in $\bar{E}$. As $\widetilde{G}_2$ is minimally persistent, and hence by Theorem 2.3 minimally rigid, by Laman's theorem, $|\widetilde{E}_2| = 2|V_2| - 3$, and as no node of V_1 is either the leader or the first follower of G, $|E_{21}| + |E_1| = 2|V_1|$. Thus

$$|\bar{E}| = |\widetilde{E}_2| + |E_{21}| + |E_1| = 2|V_2| - 3 + 2|V_1| = 2|\bar{V}| - 3.$$

Further, again from Theorem 2.3, no vertex in $\bar{G}$ has more than two outgoing edges. Also by construction, $\bar{G}$ has a leader-first follower structure, with the leader and first follower the same as that for $\widetilde{G}_2$. Thus, in view of Theorem 2.3, to complete the proof of minimal persistence, we must demonstrate that $\bar{G}$ is rigid. Thus we must show that for any $\hat{V} \subset \bar{V}$, the subgraph of $\bar{G}$ induced by $\hat{V}$ has no more than $2|\hat{V}| - 3$ edges. Choose

$$\hat{V} = \hat{V}_1 \cup \hat{V}_2 \text{ with } \hat{V}_i \subset V_i.$$

We shall count the edges by counting outgoing edges. Suppose the nodes of $\hat{V}_1$ have m outgoing edges to the nodes of $\hat{V}_2$. Then observe that the number of edges in the graph induced by $\hat{V}_1$ is no greater than $2|\hat{V}_1| - m$. Finally, to count the outgoing edges associated with $\hat{V}_2$, note that by construction, V_2 has no outgoing edges to V_1 in $\bar{G}$, and hence all outgoing edges from $\hat{V}_2$ vertices must be edges of $\widetilde{G}_2$. Thus, as $\widetilde{G}_2$ is minimally persistent from Laman's theorem, the number of outgoing edges from $\hat{V}_2$ vertices is no greater than $2|\hat{V}_2| - 3$. Hence the number of edges in the graph induced by $\hat{V}$ is no greater than

$$m + 2|\hat{V}_1| - m + 2|\hat{V}_2| - 3 = 2|\hat{V}| - 3$$

and the result holds. □

Using this lemma we will prove the following theorem.

Theorem 4.5. *Under the hypothesis of Theorem 4.1, $R(V_1)$ is generically nonsingular for every $V_1 \subset V'$.*

Proof. From Lemma 4.3 the result holds when $V_1 = V'$. Thus suppose $V_1 \neq V'$. Then we can argue that there are at least three outgoing edges from V_1 to $V - V_1$. If $g_1 = (V_1, E_1)$ is the induced subgraph, then by Laman's theorem, $|E_1| \leq 2|V_1| - 3$. Further, as V_1 does not contain either the leader or the first follower, every node in

V_1 has exactly two outgoing edges in G. Thus there must be at least three outgoing edges from V_1 to $V - V_1$.

Adopt now the notation of Lemma 4.4. Clearly $|V_2| \neq 0$. Suppose now, to obtain a contradiction, that $|V_2| = 1$. Then in the subgraph induced by $V_1 \cup V_2$ there are at least $2|V_1|$ edges, while $|V_1 \cup V_2| = |V_1| + 1$. This violates Laman's theorem as

$$2|V_1| > 2(|V_1| + 1) - 3.$$

Thus $|V_2| \geq 2$ and the conditions of Lemma 4.4 apply.

Using the notation and construction used in Lemma 4.4, call $\hat{\bar{R}}$ and $\hat{R}_2$ the matrices obtained by removing from the rigidity matrices of $\bar{G}$ and $\widetilde{G}_2$, respectively, the two columns corresponding to the common leader and one column corresponding to the common follower. By Lemma 4.4, both $\bar{G}$ and $\widetilde{G}_2$ are minimally persistent, and so by Lemma 4.2, $\hat{\bar{R}}$ and $\hat{R}_2$ are both generically nonsingular. Then the fact that no node in V_2 has an outgoing edge to the nodes of V_1 in $\bar{G}$, and that all the outgoing edges of V_1 in G are retained in $\bar{G}$, ensures that with $\times$ a *don't care* block,

$$\hat{\bar{R}} = \begin{bmatrix} R(V_1) & \times \\ 0 & \hat{R}_2 \end{bmatrix}.$$

Thus the result follows. □

Remark 4.1. Because of Theorem 4.5, we can only conclude at this point that every *even order* principal (rather than every leading principal) minor is nonzero.

We will now show that there is an ordering of vertices possible, and an ordering of the two outgoing rows associated with each vertex such that after this reordering, all leading principal minors of $\hat{R}$ are generically nonzero. We begin with the vertex reordering. Observe first the following fact, which will deal with the vertex reordering.

LEMMA 4.6. *Under the hypothesis of Theorem* 4.1, *there exists a sequence of nodes* $i_1, \dots, i_j, \dots, i_{n-2}$, *all in* V', *that has the following property: for all* $1 < j \leq n - 2$, i_j *has at most one outgoing edge in the subgraph of* G *induced by* $\{i_1, \dots, i_j\}$.

Proof. Equivalently, one must show that the node i_j has at least one edge connecting to $\{i_{j+1}, i_{j+2}, \dots, i_{n-2}, n-1, n\}$ in the original graph G. Because G is minimally persistent, by Theorem 2.3, the subgraph induced by V' has nodes which in the original graph G are the beginning vertex for at least three edges going outside the subgraph. This is because each vertex necessarily has two outgoing edges, meaning that in total there are three more than the induced subgraph is permitted to contain by Laman's theorem. Take node i_{n-2} to be the start node for any one of these edges.

Now consider the subgraph induced by the nodes in $V' - \{i_{n-2}\}$. Note that this subgraph has nodes which in the original graph G are the beginning vertices for at least three edges going outside the subgraph. Take node i_{n-3} to be the start node for any one of these edges. This procedure continues until all but two nodes have been assigned. Their assignment is trivial. □

Proof of Theorem 4.1. Select the ordering of the vertices guaranteed by Lemma 4.6. Now without loss of generality label $i_k = k$. Rows and columns $2j - 1$ and $2j$ are associated with vertex j. In the ordering of rows of $R(V')$ select the penultimate row of $R(\{1, \dots, j\})$ to be the outgoing edge of j that is not in the subgraph of G induced by $\{1, \dots, j\}$. Consider now the $(2j - 1)$th leading principal minor; i.e., under the relabeling above,

$$\det\left(\begin{bmatrix} R(\{1, \dots, j-1\}) & \times \\ 0 & x_j - x_l \end{bmatrix}\right),$$

where × is a *don't care* vector and l is the node not in $\{1,\ldots,j-1\}$ to which j's second outgoing edge goes. Then, as by Theorem 4.5 $R(\{1,\ldots,j-1\})$ is nonsingular, this determinant is nonzero. Thus after the vertex reordering and ordering of outgoing edges at each vertex, all leading principal minors of $\hat{R}$ are generically nonzero.

Remark 4.2. We remark that our control law is distributed, but the design of the control laws of the agents requires a centralized view of the particular minimally persistent formation in question. There is in fact no agreed standard definition of decentralization for a multi-agent system. One needs to distinguish a *decentralized (local) controller* from a *decentralized controller design*. The key criterion for the former is that each agent has to be autonomous and local, such that it makes decisions entirely based on local sensing and without global knowledge. The latter, however, requires that the design process of the control laws be decentralized, which is to an extent counterintuitive. One might argue that, without knowing the desirable shape of the entire formation and the agent's role or relative position in this formation, there is no way that one could assign appropriate control laws to this agent such that it knows where to be and what to do.

5. Eigenstructure analysis and an example. In this section we explain in greater depth the need for premultiplication of the update kernel by a block diagonal matrix, with 2 × 2 diagonal blocks. In particular we ask: what happens if one selects the $\hat{K}$ in (4.3) as the identity matrix? In this case, stability would require that $\hat{R}_e^{-1}\hat{R}$ have all eigenvalues in the left half plane. The basic premise of this section is that while for acyclic graphs, these conditions always hold, for graphs with cycles there may be node coordinates that result in their violation. We present an example of such instability and show how a suitably selected $\hat{K}$ repairs this instability.

In subsection 5.1 we explore the structure of $\hat{R}_e^{-1}\hat{R}$. In subsection 5.2 we show that for acyclic graphs, all eigenvalues of $\hat{R}_e^{-1}\hat{R}$ are -1. Additionally, we provide certain conditions under which several eigenvalues of $\hat{R}_e^{-1}\hat{R}$ are -1 even in graphs with cycles. Subsection 5.3 considers graphs that may have nonoverlapping cycles, and subsection 5.4 uses the eigenstructure thus established to present the example noted above.

5.1. Structure of $\boldsymbol{\hat{R}_e^{-1}\hat{R}}$. Consider any set $V_j = \{1,\ldots,j\} \subset V' = \{1,\ldots,n-2\}$, and recall the definition of $R(V_j)$ and the ordering enforced on the first $n-2$ rows of $\hat{R}$ presented in section 4. With the 2×2 matrices B_i defined in (4.2), further define

$$S(V_j) = \left(\bigoplus_{i=1}^{j} B_i\right)^{-1} R(V_j). \tag{5.1}$$

Observe that, with × denoting a *don't care* block element,

$$\hat{R}_e^{-1}\hat{R} = \begin{bmatrix} S(V') & \times \\ 0 & -1 \end{bmatrix}. \tag{5.2}$$

Thus we have the following obvious fact.

FACT 1. *At least one eigenvalue of* $\hat{R}_e^{-1}\hat{R}$ *is* -1.

Thus the interesting eigenvalues are those of $S(V')$. Further, $S(V')$ has the structure outlined below.

Define $e_1 = [1,0]'$ and $e_2 = [0,1]'$. Partition $S(V_j)$ into 2 × 2 blocks; then each matrix on the block diagonal is $-I_2$. Call X_{lr} the lrth off diagonal block element of $S(V_j)$. Then X_{lr} is nonzero if and only if there is an outgoing edge from l to r in the graph induced by V_j. There are thus at most two off diagonal nonzero block elements in each block row. If l has an outgoing edge to a node r in the subgraph induced by V_j, then if this edge information were in the $(2l-1)$th row of R,

$$X_{lr} = B_l^{-1} e_1 e_1' B_l. \tag{5.3}$$

If l has a second outgoing edge to a node s in the subgraph induced by V_j, and as this edge information must then be in the $2l$th row of R, then

$$X_{ls} = I - X_{lr} = B_l^{-1} e_2 e_2' B_l. \tag{5.4}$$

We will call X_{lr} the *edge weight* of the outgoing edge from l to r. Each edge weight has the following properties:

(a) It has rank 1.
(b) Its trace is 1.

5.2. Acyclic graphs. In this section we are concerned primarily with acyclic graphs. We first present a somewhat more general result.

THEOREM 5.1. *Suppose that q that vertices in the graph induced by $V' = \{1, 2, 3, \ldots, n-2\}$ defined above have no incoming edges and that m vertices have only one incoming edge each. Then there are at least $2q + m$ eigenvalues of $S(V')$ that are -1.*

Proof. Equivalently we need to show that $S(V') + I$ has at least $2q + m$ eigenvalues that are zero. This follows by noting that the first $2q$ columns of $S(V') + I$ are zero and the next m block columns of size $2(n-2) \times 2$ are each of rank 1. □

We next turn to graphs that are acyclic.

THEOREM 5.2. *Suppose the graph induced by V' defined above is acyclic. Then all eigenvalues of $S(V')$ are -1.*

Proof. Since the graph induced by V' is acyclic, there exists a sequence of nodes $i_1, \ldots, i_{n-2}$ such that for each $j > 1$, i_j has no outgoing edges to $\{i_1, \ldots, i_{j-1}\}$ in the graph induced by V'. Consequently under a symmetric permutation of its rows and columns $S(V')$ is upper triangular with diagonal elements all -1. This proves the result. □

COROLLARY 1. *Suppose the graph induced by V' defined above is acyclic. Then all eigenvalues of $\hat{R}_e^{-1}\hat{R}$ are -1. Therefore, with $K_i = I_2$ and $k_{n-1} = 1$, the linearized system* (3.19) *is asymptotically stable (and triangularly coupled).*

5.3. Graphs with nonoverlapping cycles. In what follows we call a graph $G'' = (V'', E'')$ a *pure cycle* if with $V'' = \{1, \ldots, k\}$, then $E'' = \{\{1, 2\}, \{2, 3\}, \ldots, \{k-1, k\}, \{k, 1\}\}$, where $\{i, j\}$ denotes an edge from i to j. If G'' is actually an induced subgraph of G, then we define its *cycle weight* to be the rank-1 matrix

$$X_{12} X_{23} \cdots X_{k-1,k} X_{k1}. \tag{5.5}$$

Again we recall that in Figure 4, $\{1, 2, 3\}$ constitutes a pure cycle, and because it is an induced subgraph, it will also have a cycle weight. We call the graph induced by V' *one with nonoverlapping cycles* if

$$V' = \bigcup_{i=1}^{r} V^i,$$

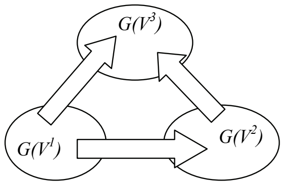


FIG. 6. *An abstracted formation showing a block of edges for a graph with pure cycles. Each $G(V^i)$ is a cycle or acyclic.*

where the graph induced by each V^i is either acyclic or a pure cycle, at least one such graph is a pure cycle, and no node of V^j has an outgoing edge to any node in

$$\bigcup_{i=1}^{j-1} V^i.$$

Then it is clear that for a graph with nonoverlapping cycles, such as in Figure 6, under a symmetric permutation of rows and columns, $S(V')$ has a block triangular structure, with $S(V^i)$ the diagonal blocks. Thus the set of eigenvalues of $S(V')$ is simply the union of the set of eigenvalues of these $S(V^i)$. In view of the results of subsection 5.2, we thus can just focus on one such V^i for which the induced subgraph is a pure cycle. Then we have the following result.

THEOREM 5.3. *Suppose the subgraph induced by $V'' = \{1, \dots, k\} \subset V'$ is a pure cycle. Define α to be the trace of the cycle weight. Then k eigenvalues of $S(V'')$ are at -1, and the remaining k are*

$$-1 + \alpha^{1/k} e^{j2\pi l/k}, \qquad l \in \{0, \dots, k-1\}. \tag{5.6}$$

Proof. Observe

$$F = I + S(V'') = \begin{bmatrix} 0 & X_{12} & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & X_{23} & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 & 0 & X_{k-1,k} \\ X_{k1} & 0 & \cdots & \cdots & \cdots & 0 & 0 \end{bmatrix}. \tag{5.7}$$

As the X_{ij} have rank 1, F has at least k zero eigenvalues. Consequently, $S(V'')$ has k eigenvalues equal to -1. For $i \in \{1, \dots, k-1\}$ call $X_{i,i+1} = a_i b_i'$ and $X_{k1} = a_k b_k'$, with a_i, b_i 2-vectors. Then

$$\alpha = b_k' a_1 \prod_{i=1}^{k-1} b_i' a_{i+1}. \tag{5.8}$$

Define

$$\gamma_1 = 1 \text{ and } \gamma_i = \frac{\alpha^{(i-1)/k}}{\prod_{j=1}^{i-1} b_j' a_{j+1}}, \quad i \in \{2, \dots, k\}, \quad \Gamma = \bigoplus_{i=1}^{k} (\gamma_i I_2), \tag{5.9}$$

for $l \in \{0, \ldots, k-1\}$,

$$W_l = \bigoplus_{i=1}^{k} \left(e^{j2\pi l(i-1)/k} I_2 \right),$$

and $\eta = [a_1', \ldots, a_k']'$. Then we assert that for each $l \in \{0, \ldots, k-1\}$,

$$FW_l\Gamma\eta = \alpha^{1/k} e^{j2\pi l/k} W_l \Gamma \eta. \tag{5.10}$$

Indeed, observe from (5.9) that for $i \in \{2, \ldots, k\}$, $\gamma_i b_{i-1}' a_i = \alpha^{1/k}\gamma_{i-1}$, and from (5.8) and (5.9) that

$$\gamma_1 b_k' a_1 = \frac{\alpha}{\prod_{j=1}^{k-1} b_j' a_{j+1}} = \alpha^{1/k}\gamma_k.$$

Thus (5.10) follows because

$$FW_l\Gamma\eta = \begin{bmatrix} e^{j2\pi l/k}\gamma_2(b_1'a_2)a_1 \\ e^{j4\pi l/k}\gamma_3(b_2'a_3)a_2 \\ \vdots \\ e^{j2\pi l(k-1)/k}\gamma_k(b_{k-1}'a_k)a_{k-1} \\ \gamma_1(b_k'a_1)a_k \end{bmatrix} = e^{j2\pi l/k} W_l \begin{bmatrix} \gamma_2(b_1'a_2)a_1 \\ \gamma_3(b_2'a_3)a_2 \\ \vdots \\ \gamma_k(b_{k-1}'a_k)a_{k-1} \\ \gamma_1(b_k'a_1)a_k \end{bmatrix}.$$

Consequently, from (5.10) the result follows. □

Observe that the theorem characterizes the eigenvectors as well. As a result of the theorem, the eigenvalues of $\hat{R}_e^{-1}\hat{R}$ contributed by a k-node nonoverlapping cycle take the form

$$-1 + \alpha^{1/k} e^{j2\pi l/k}, \quad l \in \{0, \ldots, k-1\}.$$

As we show by example in the next subsection, for suitably selected node coordinates such eigenvalues may have positive real parts.

5.4. An example. Using the results of the previous subsection we provide an example that, for suitably placed vertex positions, leads to an unstable $\hat{R}_e^{-1}\hat{R}$, implying that K_i, k_{n-1}, other than the identity, must be chosen for there to be stability. Indeed, consider the graph in Figure 4 in subsection 2.2. Note in particular that in the example the subgraphs induced by {1, 2, 3} and {4, 5, 6} are a pure cycle and acyclic; respectively, 5 is the first follower and 6 is the leader. Further, none among {4, 5, 6} has an outgoing edge to any of {1, 2, 3}. Thus indeed this graph is one we have earlier characterized to be a graph with nonoverlapping cycles. Then $\hat{R}_e^{-1}\hat{R}$ has six eigenvalues at -1 and the remaining three are at

$$-1 + \alpha^{1/3} e^{j2\pi k/3}, \quad k \in \{0, 1, 2\}. \tag{5.11}$$

Observe in this case that, the trace of the cycle weight of the cycle $\{1, 2, 3\}$ is

$$\alpha = \frac{(x_{31}y_{14} - x_{14}y_{31})(x_{12}y_{25} - x_{25}y_{12})(x_{23}y_{36} - x_{36}y_{23})}{(x_{31}y_{36} - x_{36}y_{31})(x_{12}y_{14} - x_{14}y_{12})(x_{23}y_{25} - x_{25}y_{23})},$$

where $x_{ij} = x_i - x_j$ and $y_{ij} = y_i - y_j$, and $i, j \in \{1, 2, 3, 4, 5, 6\}$.

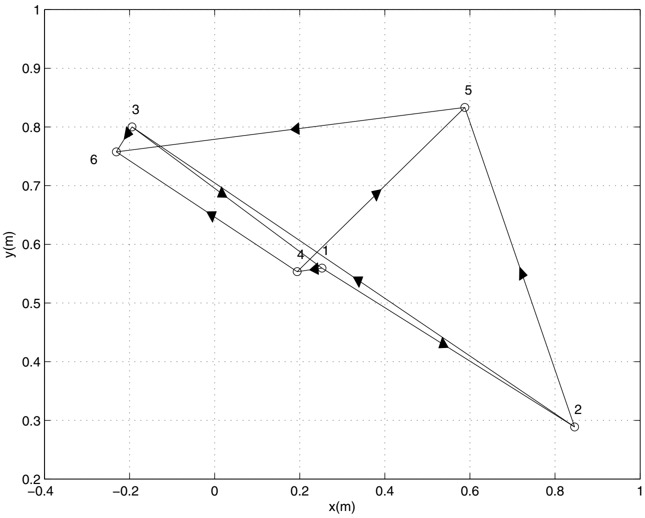


FIG. 7. *Initial setting before agent* 6 *is disturbed.*

Note that α can be alternatively expressed using angles

$$\alpha = \left(\frac{\sin \angle_{314}}{\sin \angle_{214}}\right)\left(\frac{\sin \angle_{125}}{\sin \angle_{325}}\right)\left(\frac{\sin \angle_{236}}{\sin \angle_{136}}\right), \tag{5.12}$$

where $\angle_{ijk}$ is the angle subtended by edges $\{i,j\}$ and $\{j,k\}$ at agent j, $i,j,k \in \{1,2,3,4,5,6\}$. In general the cycle weight of a pure cycle is the product of ratios, with one ratio per node appearing in the cycle. The numerator of the ratio corresponding to a particular node in the cycle is the sine of the angle subtended by the incoming edge in the cycle to this node, and the outgoing edge from this node leaving the cycle. The denominator is the sine of the angle between the two outgoing edges of this node.

This expression provides some geometric clue for the formation stability, since a sine will be small if the angle approaches 0 or π, which means that a certain set of three agents is nearly collinear. Yet a large α and hence potential instability may not occur even if some nodes are near to being collinear. Thus, suppose in Figure 4 that if nodes 1, 2, and 4 are near collinear but $\angle_{314} < \angle_{214}$, then instability may well be avoided.

For an instantiation of the formation graph given in Figure 4, choose the six agent positions to be $p_1 = (0.2902, 0.5409)$, $p_2 = (0.8637, 0.2302)$, $p_3 = (-0.1388, 0.8117)$, $p_4 = (0.2316, 0.5387)$, $p_5 = (0.6438, 0.7909)$, and $p_6 = (-0.1784, 0.7716)$. An instantiation of this graph is shown in Figure 7.

Then $\alpha = 1.1407$ and (5.11) for $k = 0$ is real and positive, implying instability. Note in this particular example that α assumes an intermediate value, and none of the relevant angles appearing in either the numerator or the denominator of (5.12) is close to 0 or π. Thus, it is not relevant that agents 1, 2, and 3 are nearly collinear.

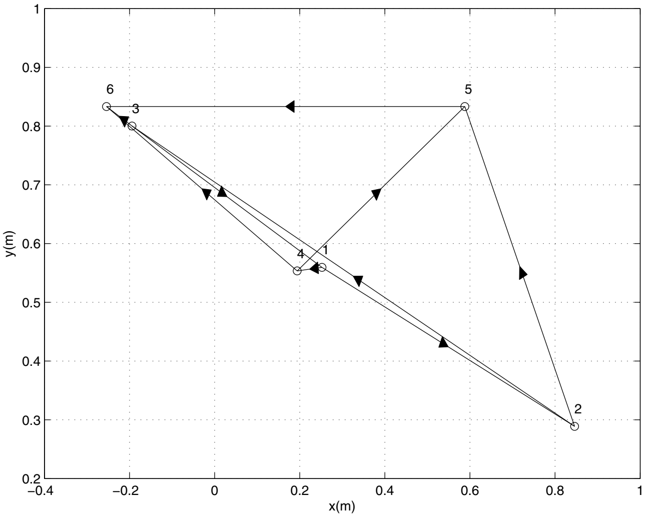


FIG. 8. *Initial setting after agent* 6 *is disturbed.*

On the other hand, it is readily checked that with the choice

$$\hat{K} = \begin{bmatrix} 5.4317 & -3.0234 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1.9306 & -0.7167 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 10.2598 & -6.9802 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.2448 & 4.0662 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0.4261 & 0.2535 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0.0634 & 0.1032 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

the eigenvalues of $\hat{K}\hat{R}_e^{-1}\hat{R}$ are $\{-0.0515 - 0.0473i,\ -0.0515 + 0.0473i,\ -0.4298,\ -1.0000,\ -1.0000,\ -1.0000,\ -3.2276,\ -4.7944,\ -10.1634\}$. Observe that this $\hat{K}$ has the structure imposed by (3.17) and provides a stable solution.

We next perturb the coordinates in the graph of Figure 7 to that in Figure 8. In the latter figure, agent 6 moves to $p_6 = (-0.1829, 0.7909)$. The nonlinear control law of (3.5), (3.7) is applied, with K_i as in the previous subsection. The simulation results are shown in Figures 9 and 10 and indicate rapid convergence.

6. Conclusion. This paper has only started on what is likely to be a fairly long road, developing efficacious control designs for maintaining formation shape. The methods of this paper do little more than demonstrate stabilizability. The particular stability theorem we are relying on, involving multiplying a matrix with nonzero leading principal minors by a diagonal matrix to make it stable, is almost certainly novel; however, it does not address the achieving of other control objectives apart from stability. In fact there is a broader list of issues that need to be addressed in the future, and we record some as follows:

(a) The control laws of this paper should really be regarded as nonlinear laws, with the rigidity matrix varying in the course of the motion. We have assumed small motions in order to justify an analysis using a linearized system. An

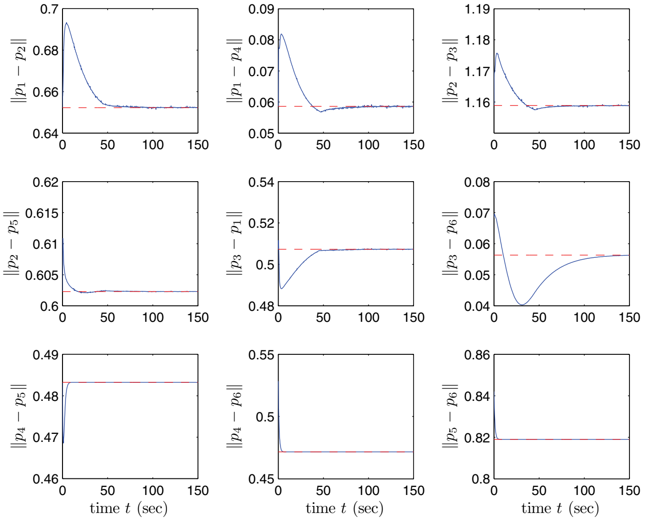


FIG. 9. *Distance keeping during the shape regulation process.*

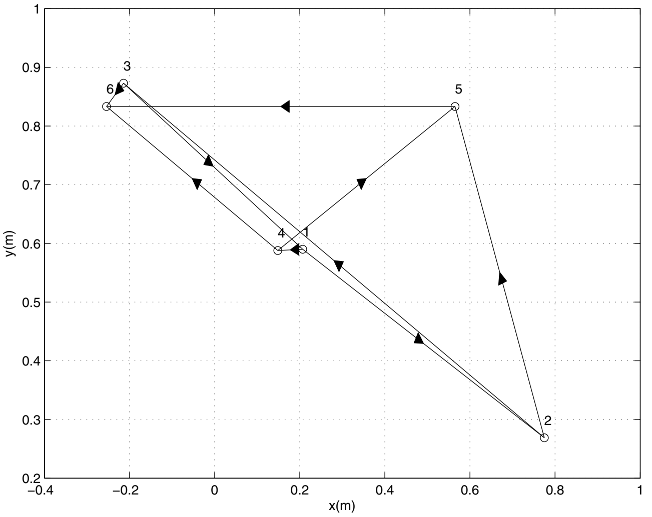


FIG. 10. *Final setting after shape regulation, i.e., at time $t = 150$ sec.*

immediate task would be to demonstrate stability of the nonlinear algorithm for a sizeable domain of attraction, and at the same time, or separately, construct a framework that embraces all minimally persistent formations, and not just ones of the leader-first follower type.

(b) We could have chosen different variables in which to describe the problem; for example, in another work [22], the dependent variables in the key differential equations were actual edge lengths rather than coordinate values. From a linearized point of view, it turns out that this makes no difference, but working with edge lengths may allow for a clearer understanding of the nonlinear behavior under big perturbations. Certainly in working with edge lengths in [22], we were able to obtain results that were almost global in their applicability.

(c) It may be important to study topologically balanced graphs (and then perhaps graphs with weights which can balance them) [11, 17] and see whether the results can be obtained far more easily. This is because for directed graphs and the consensus (flocking) problem, balanced graphs allow an easy solution [11], bringing the Laplacian into the picture; there is no guarantee that this will be so for the shape maintenance problem, which certainly needs to be viewed as a nonlinear problem, but it might be so. Note, though, that having a balanced graph precludes having a leader-follower structure, although it might be possible to separate these two vertices, much as we have done in this paper. To the extent that we are using weights here, which are the entries of the diagonal scaling matrix Λ, one might even wonder whether these could be put into a weighted balancing framework.

(d) Three-dimensional formations will not necessarily be a straightforward generalization, since the persistence concept is more difficult to generalize than might at first appear. Further, there is lacking a full graphical characterization of rigidity in three dimensions analogous to Laman's theorem, which we drew on heavily in this paper. On the other hand, it is reasonable to conjecture that the critical technical requirement, that a certain submatrix of the rigidity matrix will have all leading principal minors nonzero, will continue to hold.

(e) There are only limited insights here into which structures are difficult to control, in the sense that very large control signals will need to be used or noise will be a great problem. Such insights can be obtained through a more detailed investigation of formulas such as (5.12).

(f) Next, it is quite evident that it is desirable to include redundancy in formations to allow for loss of communication links or loss of an agent. This means we need to be able to handle nonminimally persistent problems, especially where there is noise in the measurements; for there is not then an obvious single point toward which a follower agent should aim.

These remarks of course do not exhaust the problems. One could imagine treating agents with mass, inertia, orientation; other control laws; maintenance of shape mixed with formation motion objectives, including obstacle avoidance; minimization of control energy; agents which move asynchronously; or formations with communications delays. These are all examples of issues which are relevant and have yet to be addressed.

REFERENCES

[1] J. Baillieul and A. Suri, *Information patterns and hedging Brockett's theorem in controlling vehicle formations*, in Proceedings of the 42nd IEEE Conference on Decision and Control, Maui, HI, 2003, pp. 556–563.

[2] R. Olfati-Saber and R. M. Murray, *Distributed cooperative control of multiple vehicle formations using structural potential functions*, in Proceedings of the 15th IFAC World Congress, Barcelona, Spain, 2002, pp. 1–7.

[3] R. Olfati-Saber and R. M. Murray, *Distributed structural stabilization and tracking for formations of dynamic multi-agents*, in Proceedings of the 41st IEEE Conference on Decision and Control, Las Vegas, NV, 2002, pp. 209–215.

[4] R. Olfati-Saber and R. M. Murray, *Graph rigidity and distributed formation stabilization of multivehicle systems*, in Proceedings of the 41st IEEE Conference on Decision and Control, Las Vegas, NV, 2002, pp. 2965–2971.

[5] R. Olfati-Saber, W. B. Dunbar, and R. M. Murray, *Cooperative control of multi-vehicle systems using cost graphs and optimization*, in Proceedings of the American Control Conference, Denver, CO, 2003, pp. 2217–2222.

[6] H. G. Tanner, A. Jadbabaie, and G. J. Pappas, *Stable flocking of mobile agents, part* I*: Fixed topology*, in Proceedings of the 42nd IEEE Conference on Decision and Control, Maui, HI, 2003, pp. 2010–2015.

[7] H. G. Tanner, A. Jadbabaie, and G. J. Pappas, *Stable flocking of mobile agents, part* II*: Dynamic topology*, in Proceedings of the 42nd IEEE Conference on Decision and Control, Maui, HI, 2003, pp. 2016–2021.

[8] P. Tabuada, G. J. Pappas, and P. Lima, *Cyclic directed formations of multi-agent systems*, in Proceedings of the 40th IEEE Conference on Decision and Control, Arlington, VA, 2001, pp. 56–61.

[9] J. M. Hendrickx, B. D. O. Anderson, J.-C. Delvenne, and V. D. Blondel, *Directed graphs for the analysis of rigidity and persistence in autonomous agent systems*, Internat. J. Robust Nonlinear Control, 17 (2007), pp. 960–981.

[10] C. Yu, J. M. Hendrickx, B. Fidan, B. D. O. Anderson, and V. D. Blondel, *Three and higher dimensional autonomous formations: Rigidity, persistence and structural persistence*, Automatica, 43 (2007), pp. 387–402.

[11] D. Lee and M. Spong, *Stable flocking of inertial agents on balanced communication graphs*, in Proceedings of the American Control Conference, Minneapolis, MN, 2006, pp. 2136–2141.

[12] Z. Lin, M. E. Broucke, and B. A. Francis, *Local control strategies for groups of mobile autonomous agents*, IEEE Trans. Automat. Control, 49 (2004), pp. 622–629.

[13] J. A. Marshall, M. E. Broucke, and B. A. Francis, *Formations of vehicles in cyclic pursuit*, IEEE Trans. Automat. Control, 49 (2004), pp. 1963–1974.

[14] R. Olfati-Saber, *Flocking for multi-agent dynamic systems: Algorithms and theory*, IEEE Trans. Automat. Control, 51 (2006), pp. 410–420.

[15] W. Ren and R. W. Beard, *A decentralized scheme for spacecraft formation flying via the virtual structure approach*, J. Guidance Control Dynam., 27 (2004), pp. 73–82.

[16] F. E. Schneider and D. Wildermuth, *Experimental comparison of a directed and a non-directed potential field approach to formation navigation*, in Proceedings of IEEE International Symposium on Computational Intelligence in Robotics and Automation, Espoo, Finland, 2005, pp. 21–26.

[17] J. Cortes, *Distributed algorithms for reaching consensus on general functions*, Automatica, 44 (2008), pp. 726–737.

[18] T. Tay and W. Whiteley, *Generating isostatic frameworks*, Structural Topology, No. 11, 1985, pp. 21–69.

[19] G. Laman, *On graphs and rigidity of plane skeletal structures*, J. Engrg. Math., 4 (1970), pp. 331–340.

[20] B. Fidan, C. Yu, and B. D. O. Anderson, *Acquiring and maintaining persistence of autonomous multi-vehicle formations*, IET Control Theory Appl., 1 (2007), pp. 452–460.

[21] S. L. Smith, M. E. Broucke, and B. A. Francis, *Stabilizing a multi-agent system to an equilibrium polygon formation*, in Proceedings of the 17th International Symposium on Mathematical Theory of Networks and Systems, Kyoto, Japan, 2006, pp. 2415–2424.

[22] B. D. O. Anderson, C. Yu, S. Dasgupta, and A. S. Morse, *Control of a three coleaders persistent formation in the plane*, Systems Control Lett., 56 (2007), pp. 573–578.

[23] M. Cao, A. S. Morse, C. Yu, B. D. O. Anderson, and S. Dasgupta, *Controlling a triangular formation of mobile autonomous agents*, in Proceedings of the 46th IEEE Conference on Decision and Control, New Orleans, MA, 2007, pp. 3603–3608.

[24] R. S. Johnson, *Singular Perturbation Theory. Mathematical and Analytical Techniques with Applications to Engineering*, Springer, New York, 2005.

[25] S. Friedland, *On inverse multiplicative eigenvalue problems for matrices*, Linear Algebra Appl., 12 (1985), pp. 127–137.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 1, pp. 234–265

© 2009 Society for Industrial and Applied Mathematics

COORDINATED PATH-FOLLOWING IN THE PRESENCE OF COMMUNICATION LOSSES AND TIME DELAYS*

R. GHABCHELOO†, A. P. AGUIAR†, A. PASCOAL†, C. SILVESTRE†, I. KAMINER‡, AND J. HESPANHA§

Abstract. This paper addresses the problem of steering a group of vehicles along given spatial paths while holding a desired time-varying geometrical formation pattern. The solution to this problem, henceforth referred to as the coordinated path-following (CPF) problem, unfolds in two basic steps. First, a path-following (PF) control law is designed to drive each vehicle to its assigned path, with a nominal speed profile that may be path dependent. This is done by making each vehicle approach a virtual target that moves along the path according to a conveniently defined dynamic law. In the second step, the speeds of the virtual targets (also called coordination states) are adjusted about their nominal values so as to synchronize their positions and achieve, indirectly, vehicle coordination. In the problem formulation, it is explicitly considered that each vehicle transmits its coordination state to a subset of the other vehicles only, as determined by the communications topology adopted. It is shown that the system that is obtained by putting together the PF and coordination subsystems can be naturally viewed as either the feedback or the cascade connection of the latter two. Using this fact and recent results from nonlinear systems and graph theory, conditions are derived under which the PF and the coordination errors are driven to a neighborhood of zero in the presence of communication losses and time delays. Two different situations are considered. The first captures the case where the communication graph is alternately connected and disconnected (brief connectivity losses). The second reflects an operational scenario where the union of the communication graphs over uniform intervals of time remains connected (uniformly connected in mean). To better root the paper in a nontrivial design example, a CPF algorithm is derived for multiple underactuated autonomous underwater vehicles (AUVs). Simulation results are presented and discussed.

Key words. coordination control, communication losses and time delays, path-following, autonomous underwater vehicles

AMS subject classifications. 93A14, 93D15, 93C85, 93C10, 93A13

DOI. 10.1137/060678993

1. Introduction.

Increasingly challenging mission scenarios and the advent of powerful embedded systems, sensors, and communication networks have spawned widespread interest in the problem of coordinated motion control of multiple autonomous vehicles. The types of applications considered are numerous and include aircraft and spacecraft formation control [6], [23], [29], [36], [39], [49], [50], [60], [30], [4], [64], coordinated control of land robots [16], [52], [22], [58], control of multiple surface and underwater vehicles [17], [26], [34], [63], [13], and networked control of robotic systems

*Received by the editors December 31, 2006; accepted for publication (in revised form) November 7, 2008; published electronically February 11, 2009. This research is supported in part by project GREX/CEC-IST under contract 035223, project NAV-Control/FCT-PT (PTDC/EEAACR/65996/2006), the FREESUBNET RTN of the CEC, the FCT-ISR/IST pluriannual funding program (through the POS C initiative in cooperation with FEDER), and by NSF grant ECS-0242798. The first author benefitted from a Ph.D. scholarship of FCT.

http://www.siam.org/journals/sicon/48-1/67899.html

†Institute for Systems and Robotics and the Department of Electrical Engineering and Computers, Instituto Superior Técnico, Av. Rovisco Pais, 1, 1049-001 Lisboa, Portugal (reza@isr.ist.utl.pt, pedro@isr.ist.utl.pt, antonio@isr.ist.utl.pt, cjs@isr.ist.utl.pt).

‡Department of Mechanical and Astronautical Engineering, Naval Postgraduate School, Monterey, CA 93943 (kaminer@nps.navy.edu).

§Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106-9560 (hespanha@ece.ucsb.edu).

[15], [14], [33], [41], [37], [43], [48]. To meet the requirements imposed by these and related applications, a new control paradigm is needed that must necessarily depart from classical centralized control strategies. Centralized controllers deal with systems in which a single controller possesses all the information required to achieve desired control objectives, including stability and performance requirements. In many of the applications envisioned, however, the highly distributed nature of the vehicles' sensing and actuation modules and the constraints imposed by the intervehicle communication network make it impossible to tackle the problems in the framework of centralized control theory. In part due to these reasons, there has been over the past few years a flurry of activity in the area of multiagent networks with application to engineering and science problems. The list of related research topics is vast and includes parallel and distributed computing [7], distributed decision making [61], synchronization in oscillator networks [46], flocking of mobile autonomous agents [5], [18], [28], [54], state agreement and consensus problems [20], [38], [51], [40], [44], [42], [11], asynchronous consensus protocols [9], [61], graph theory and graph connectivity [45], [56], [32], rigidity and persistence in autonomous formations [62], adaptive and distributed coordination algorithms for mobile sensing networks [12], [11], and concurrent synchronization in dynamic system networks [48]. See also [55] and the references therein for general expositions on large-scale dynamical systems and decentralized control of complex systems that bear affinity with the issues addressed in this paper.

In spite of significant progress made in these areas, much work remains to be done to develop strategies capable of yielding robust performance of a fleet of vehicles in the presence of complex vehicle dynamics, severe communication constraints, and partial vehicle failures. These difficulties are especially challenging in the field of marine robotics for two main reasons: (i) often, the dynamics of marine vehicles cannot be greatly simplified for control design purposes; and (ii) underwater communications and positioning rely heavily on acoustic systems, which are plagued with intermittent failures, latency, and multipath effects.

Inspired by recent theoretical and practical developments in the areas of multiple vehicle control, we consider the problem of *coordinated path-following (CPF) control*, where *multiple vehicles are required to follow prespecified paths while keeping a desired, possibly time-varying, geometric formation pattern.* These objectives must be met in the presence of communication losses and delays. The problem arises, for example, in the operation of multiple autonomous underwater vehicles (AUVs) for fast acoustic coverage of the seabed [47]. In this application, two or more vehicles maneuver above the seabed, at either the same or different depths, along geometrically similar spatial paths and map the seabed using identical suites of acoustic sensors. By requesting that the vehicles move along the paths so that the projections of the acoustic beams on the seabed have a certain degree of overlapping, large areas can be mapped in a short time. These objectives impose strict constraints on the vehicle formation pattern.

A number of other scenarios can be envisioned that require CPF control of marine vehicles. Examples include underwater vehicle formation control for 3D vision-based marine habitat mapping, ship underway replenishment [34], and missions where temporal and spatial path deconfliction are critical [30]. Similar problems arise in the area of air vehicle control. All of these scenarios share the requirements that a number of vehicles maneuver along predetermined paths, at nominal speed profiles that may be path dependent, and keep a possibly time-varying formation pattern. Absolute time requirements are not part of the problem. As such, they depart considerably from other related problems such as vehicle rendezvous maneuvers and swarm formation control. The manner in which the paths and the formation are planned depend on the specific problem at hand, for example, using a time optimality criterion when fast coordinated maneuvering between initial and final positions is required, minimizing an energy-related criterion when the objective is to scan a certain area or volume densely and energy is at a premium, or using a combination of criteria that include geometric constraints when collision avoidance is important. See, for example, [30] for the case of unmanned air vehicles.

In this paper we formulate and solve the problem of CPF by explicitly taking into account the vehicle dynamics and the topology of the underlying communication network, subject to communication losses and delays. The reader is referred to [17], [21], [22], [35], [58] and the references therein for a historical overview of the topic and a perspective of the sequence of motion control problem formulations and solutions upon which the present work builds. See also [19], [57], [47] for an in-depth introductory exposition to the topic at hand. For the sake of clarity, it is important to point out that in the scope of the problem at hand, PF and CPF have also been referred to as output maneuvering and synchronization of multiple maneuvering systems, respectively [58]. A comprehensive survey of related results on consensus in multivehicle cooperative control can be found in [42], [44], and [51].

The solution to the problem of CPF that we propose unfolds in two basic steps. First, a PF control law is designed to drive each vehicle to its assigned path, with a nominal speed profile that may be path dependent. This is done by making each vehicle approach a virtual target that moves along the path according to a conveniently defined dynamic law. Each vehicle has access to a set of local measurements only. In the second step, the speeds of the virtual targets (also called coordination states) are adjusted about their nominal values so as to synchronize their positions and achieve, indirectly, vehicle coordination. The vehicles are allowed to exchange only limited information with their immediate neighbors. Without being too rigorous, it can be said that the strategy proposed abides by a separation principle whereby the PF and coordinated motion control designs are almost decoupled. This simplifies the overall design process. Furthermore, it has the virtue of leaving essentially to each vehicle the task of dealing with external disturbances acting upon it, directly at the PF level.

The mathematical setup adopted in the paper is well rooted in Lyapunov stability and graph theory. At the pure PF level, two types of control laws, henceforth referred to as Type I and Type II, are developed. The difference between them lies in the types of dynamics chosen for the virtual targets along the paths.

Key concepts from input-to-state stability theory [59] are also used to derive results on the *stability, performance, and robustness* of the overall system that results from putting together the PF and vehicle coordination subsystems. Here, we use the fact that combination of the above systems takes either a feedback interconnection or a cascade form, depending on whether the underlying PF laws are of Type I or Type II. The results are quite general in that they apply to a large class of PF control systems satisfying a certain input-to-state stable (ISS) property. For the sake of clarity and completeness, the paper derives a PF strategy for a class of underactuated AUVs that meets the required ISS property.

The key contribution of the paper is the study of the combined behavior of the PF and coordination systems in the presence of temporary communication losses and transmission delays. To deal with communication losses, the paper proposes two frameworks for studying the effect of communication failures and delays on the performance of the overall vehicle formation. The first framework, brief connectivity losses (BCLs), refers to the situation where the communication graph changes in time, alternating between connected and disconnected graphs. Here, we borrow from and expand previous results on systems with brief instabilities, namely, by deriving a new small-gain theorem that applies to the feedback interconnection of these systems. See [25] and the references therein for an introduction to systems with brief instabilities and their application to control systems analysis and design. The second framework, uniformly connected in mean (UCM), applies to the case where the communication graph may even fail to be connected at any instant of time; however, we assume there is a finite time $T > 0$ such that over any interval of length T the union of the different graphs is connected. This framework is motivated by the work in [37], [38], [40]. To the best of our knowledge, this is the first time that the impact of intermittent failures on coordinated PF is analyzed from a quantitative point of view and estimates on the rate of decay of all closed-loop error signals are obtained. The impact of communication delays on the overall system performance is also analyzed for the case of homogeneous delays and PF systems of Type II. Conditions are derived under which the PF errors become arbitrarily small and the cooperation errors approach zero exponentially. For related results on the consensus problems for systems with nonhomogeneous delays, see [20].

The paper is organized as follows. Section 2 formulates the PF and vehicle coordination problems and describes general stability-related properties that are met by the PF closed-loop subsystem of each vehicle. Section 3 introduces some basic notation, summarizes important results on graph theory, and develops the tools that will be used to study the different types of communication topologies considered in the paper. Section 4 derives a useful small-gain theorem for the feedback interconnection of systems with brief instabilities. Section 5 studies the CPF problem in the case where the communications network is subjected to communication losses with no time delays. Section 6 extends some of the results of section 5 to deal with switching communication networks and time delays. An illustrative example is presented in section 7, where a CPF control algorithm for a general class of underactuated AUVs is derived. The results of simulations are also described. Finally, section 8 contains the main conclusions and describes problems that warrant further research. The proofs of several statements are included in the appendix.

2. Problem statement.

This section provides a rigorous formulation of the PF and coordination problems that are the main subjects of the paper. Consider a group of n vehicles numbered $1, \dots, n$. We let the dynamics of vehicle i be modeled by a general system of the form

$$\begin{aligned} \dot{x}_i &= f_i(x_i, u_i, w_i), \\ y_i &= h_i(x_i, v_i), \end{aligned} \tag{2.1}$$

where $x_i \in \mathbb{R}^n$ is the state, $u_i \in \mathbb{R}^m$ is the control signal, and $y_i \in \mathbb{R}^q$ is the output that we require to reach and follow a path $y_{d_i}(\gamma_i) : \mathbb{R} \to \mathbb{R}^q$ parameterized by $\gamma_i \in \mathbb{R}$. Signals w_i and v_i denote the disturbance inputs and measurement noises, respectively. Later in section 7, an example will be given where the dynamics of (2.1) are those of a very general class of AUVs. In that case, the output y_i corresponds to the position of the vehicle with respect to an inertial coordinate frame.

For any continuous, differentiable timing law $\gamma_i(t)$, define the PF and speed tracking error variables

$$e_i(t) := y_i(t) - y_{d_i}(\gamma_i(t)) \tag{2.2}$$

and

$$\eta_i(t) := \dot{\gamma}_i(t) - v_{r_i}(t), \tag{2.3}$$

respectively, where $v_{r_i}(t) \in \mathbb{R}$ denotes a desired temporal speed profile to be defined.

Inspired by the work in [3, 22, 57], we start by defining the problem of PF for each vehicle. In what follows, $\|.\|$ denotes both the Euclidean norm of a vector and the spectral norm of a matrix.

Definition 2.1 (PF problem). *Consider a vehicle with dynamics* (2.1)*, together with a spatial path $y_{d_i}(\gamma_i)$; $\gamma_i \in \mathbb{R}$, to be followed and a desired, predetermined temporal speed profile $v_{r_i}(t)$ to be tracked. Let the PF error and the speed tracking error be as in* (2.2) *and* (2.3)*, respectively. Given $\epsilon > 0$, design a feedback control law for u_i such that all closed-loop signals are bounded and both $\|e_i\|$ and $|\eta_i|$ converge to a neighborhood of the origin of radius ϵ.*

Stated in simple terms, the problem above amounts to requiring that the output y_i of a vehicle converge to and remain inside a tube centered around the desired path y_{d_i}, while ensuring that its rate of progression $\dot{\gamma}_i$ also converge to and remain inside a tube centered around the desired speed profile $v_{r_i}(t)$.

We assume that the PF controllers adopted meet a number of technical conditions described next. In section 7, as an example, we introduce a PF controller for a general underactuated vehicle that meets these conditions. The interested reader will find in [22], [57], and the references therein related material on PF control of nonlinear systems. See also [3] for an interesting discussion on the possible advantages of PF versus trajectory tracking control. Namely, the fact that PF control for nonminimum phase systems removes the performance limitations that are inherent to trajectory tracking schemes.

In preparation for the development that follows, we set $v_{r_i}(t) = v_L(\gamma_i(t), t) + \tilde{v}_{r_i}(t)$, where $v_L(\gamma_i, t)$ is a nominal predetermined speed profile and $\tilde{v}_{r_i}$ can be seen as a perturbation component of v_{r_i} about v_L. Later, it will be shown that $v_L(.,.)$ is common to all the vehicles and known in advance and that

$$\tilde{v}_{r_i}(t) := v_{r_i}(t) - v_L(\gamma_i(t), t) \tag{2.4}$$

(the remaining component of $v_{r_i}(t)$) is not known beforehand. We assume that $y_{d_i}(.)$ is sufficiently smooth with respect to its argument. We further assume that $v_L(.,.)$ is bounded and globally Lipschitz with respect to the first argument, that is, $\exists\, v_M, l > 0$, such that $|v_L(\gamma_i, t)| \leq v_M$ and $|v_L(\gamma_i, t) - v_L(\gamma_j, t)| \leq l|\gamma_i - \gamma_j|$.

Consider vehicle i and assume a feedback control law $u_i = u_i(x_i, y_{d_i}, v_L)$ exists that solves the PF problem of Definition 2.1. Let the corresponding closed-loop PF system be described by

$$\dot{\zeta}_i = f_{c_i}(t, \zeta_i, \tilde{v}_{r_i}, d_i), \tag{2.5}$$

where d_i subsumes all the exogenous inputs (including disturbances and measurement noises), $\tilde{v}_{r_i}$ is defined as in (2.4), and state vector ζ_i includes necessarily e_i but may or may not include η_i. Two types of PF strategies will be considered:

1. *Type* I. In this strategy, variable η_i plays the role of an auxiliary state for the PF algorithm and specifies the evolution of γ_i. In this case η_i is a state of the closed-loop PF system, that is, ζ_i includes η_i.
2. *Type* II. This strategy is equivalent to making $\eta_i = 0$. The dynamics of $\dot{\gamma}_i$ are simply $\dot{\gamma}_i = v_{r_i}$. Clearly, in this case ζ_i does not include η_i.

We now recall the definitions of input-to-state stable (ISS) and input-to-state practical stable (ISpS) for a dynamical system. See [59] and [31, p. 217] for details on ISS and ISpS and their relation to Lyapunov theory. System (2.5) is said to be

ISpS if there exist a class $\mathcal{KL}$ function $\beta(.,.)$, class $\mathcal{K}$ functions[1] $\rho_i(.)$; $i = 1, 2$, and a constant $\rho_3 > 0$ such that for any inputs $\tilde{v}_{r_i}$ and d_i and any initial condition $\zeta_i(t_0)$, the solution of (2.5) satisfies

$$\|\zeta_i(t)\| \leq \beta(\|\zeta_i(t_0)\|, t-t_0) + \rho_1\left(\sup_{t_0\leq s\leq t} \|\tilde{v}_{r_i}(s)\|\right) + \rho_2\left(\sup_{t_0\leq s\leq t} \|d_i(s)\|\right) + \rho_3 \quad \forall t \geq t_0.$$

System (2.5) is said to be ISS if it satisfies the conditions of ISpS with $\rho_3 = 0$.

Assumption 2.2. *We assume there exists a Lyapunov function $W_i(t, \zeta_i)$ for* (2.5) *satisfying*

$$\underline{\alpha}_1\|\zeta_i\|^2 \leq W_i \leq \bar{\alpha}_1\|\zeta_i\|^2, \tag{2.6}$$

$$\dot{W}_i \leq -\lambda_1 W_i + \rho_1|\tilde{v}_{r_i}|^2 + \rho_2 d_i^2, \tag{2.7}$$

where λ_1, ρ_1, ρ_2, $\underline{\alpha}_1$, and $\bar{\alpha}_1$ are positive values and $\dot{W}_i$ is computed along the solutions of (2.5)*, that is,*

$$\dot{W}_i = \frac{\partial W_i}{\partial t} + \frac{\partial W_i}{\partial \zeta_i} f_{c_i}.$$

With this assumption, the closed-loop PF system (2.5) is ISS with input $(d_i, \tilde{v}_{r_i})$ and state ζ_i. To verify this, integrate (2.7) and use (2.6) to obtain

$$\underline{\alpha}_1\|\zeta_i(t)\|^2 \leq \bar{\alpha}_1\|\zeta_i(t_0)\|^2 e^{-\lambda_1(t-t_0)} + \frac{\rho_1}{\lambda_1}\sup|\tilde{v}_{r_i}|^2 + \frac{\rho_2}{\lambda_1}\sup|d_i|^2,$$

and therefore

$$\|\zeta_i(t)\| \leq \alpha\|\zeta_i(t_0)\| e^{-0.5\lambda_1(t-t_0)} + \rho_v \sup|\tilde{v}_{r_i}| + \rho_d \sup|d_i|,$$

with $\alpha = \sqrt{\bar{\alpha}_1/\underline{\alpha}_1}$, $\rho_v = \sqrt{\rho_1/(\lambda_1\underline{\alpha}_1)}$, and $\rho_d = \rho_2/(\lambda_1\underline{\alpha}_1)$.

Assuming a PF controller has been implemented for each vehicle, it now remains to coordinate (that is, synchronize) the entire group of vehicles so as to achieve a desired formation pattern compatible with the paths adopted. As will become clear, this will be achieved by adjusting the desired speeds of the vehicles as functions of the "along-path" distances among them. To better grasp the key ideas involved in the computation of these distances, consider as a simple example the case of a fleet of vehicles that are required to move along parallel straight lines and keep themselves aligned along a direction perpendicular to the lines. See Figure 1 for the case of two vehicles.

Let Γ_i denote the desired path to be followed by vehicle i and assume Γ_i is simply parameterized by s_i, the path length. In other words, $\gamma_i = s_i$. Because each vehicle approaches the path as close as required, that is, because $y_i(t)$ becomes arbitrarily close to to $y_{d_i}(\gamma_i)$, it follows that the vehicles are (asymptotically) synchronized if $\gamma_{ij}(t) := \gamma_i(t) - \gamma_j(t) \to 0\ \forall i, j \in \mathcal{N} := \{1, \ldots, n\}$. This shows that in the case of translated straight lines $\gamma_{i,j} = s_i - s_j$ is a good measure of the along-path distances among the vehicles. Similarly, in the case of vehicles that must be aligned along the radii of nested circumferences as in Figure 2, an appropriate measure of the distances among the vehicles is angle $\gamma_i = s_i/R_i$ where s_i denotes path length and

[1] A function ρ is of class $\mathcal{K}$ if it is strictly increasing and $\rho(0) = 0$. A function $\beta(r, s)$ belongs to class $\mathcal{KL}$ if the mapping $\beta(r, s)$ is of class $\mathcal{K}$ for a fixed s, is decreasing with respect to s for a fixed r, and $\beta(r, s) \to 0$ as $t \to \infty$.

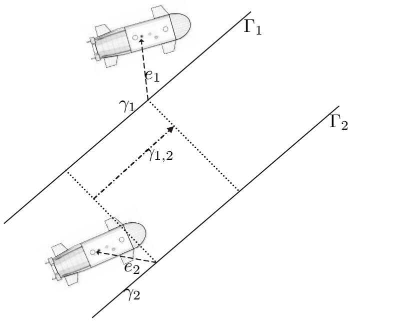


FIG. 1. *Along-path distances: straight lines.*

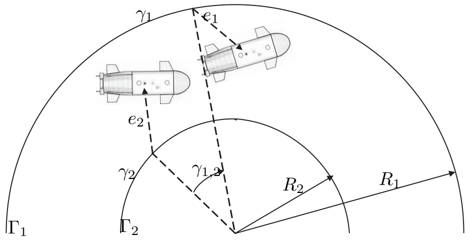


FIG. 2. *Along-path distances: circumferences.*

R_i is the radius of circumference i. Clearly, this corresponds to adopting different parameterizations for the paths that correspond to normalizing their lengths. In both cases, we say that the vehicles are coordinated if the corresponding along-path distance is zero, that is, $\gamma_i - \gamma_j = 0$. Coordination is achieved by adjusting the desired speed of each vehicle i as a function of the along-path distances γ_{ij}; $j \in N_i$, where N_i denotes the set of vehicles with which vehicle i communicates. For arbitrary types of paths and coordination patterns, an adequate choice of path parameterizations will allow for the conclusion that the vehicles are *coordinated* or, in equivalent terms, are *synchronized*/have reached *agreement*, iff $\gamma_{i,j} = 0\ \forall j, i \in \mathcal{N}$; see [22], [16]. Since the objective of the coordination is to coordinate variables γ_i, we will refer to them as *coordination states*.

We will require that the formation as a whole (group of multiple vehicles) travel at an assigned speed profile $v_L(\gamma_i, t)$ when coordinated, that is, $\dot{\gamma}_i = v_L\ \forall i \in \mathcal{N}$, where v_L is allowed to be a function of path parameter γ and time t. This follows from the fact that $v_L(\gamma_i, t) = v_L(\gamma_j, t)$ when $\gamma_i = \gamma_j$. This issue requires clarification. Note that the desired speed assignment is given in terms of the time derivatives of the coordination states γ_i, not in terms of the inertial speeds (actual time derivative of the positions) of the vehicles undergoing synchronization. In the limit, as shown later, the combined PF and coordination algorithms will ensure that the coordination states will be equal and the vehicle speeds will naturally approach $\frac{ds_i}{dt}$: $i \in \mathcal{N}$, so that $\frac{d\gamma_i}{dt} = \frac{d\gamma_i}{ds_i}\frac{ds_i}{dt} = v_L$. Thus, $\frac{ds_i}{dt} = v_L / \frac{d\gamma_i}{ds_i}$ which shows clearly how coordination states speed and inertial speeds depend on the path parameterizations adopted. In the case of the circumferences above, the latter relationship yields simply $\frac{ds_i}{dt} = R_i v_L$. Notice how the speed assignment in terms of the coordination states avoids the need to specify the actual inertial speeds of the vehicles in an inertial reference frame, which would be quite cumbersome. Instead, all that is required is to specify the speeds of the coordination states which are equal and degenerate simply into v_L.

From (2.3), the evolution of the coordination state γ_i, $i \in \mathcal{N}$, is governed by

$$\dot{\gamma}_i(t) = v_{r_i}(t) + \eta_i(t), \tag{2.8}$$

where the speed tracking errors η_i are viewed as disturbance-like input signals and the speed profiles v_{r_i} are taken as control signals that must be assigned to yield coordination of the states γ_i. To achieve this objective, information is exchanged through an intervehicle communication network. Typically, all-to-all communications are impossible to achieve. In general, $\dot{\gamma}_i$ will be a function of γ_i and of the coordination states of the so-called neighboring agents defined by set N_i. For simplicity of presentation, throughout this paper we assume that the communication links are bidirectional, that is, $i \in N_j \Leftrightarrow j \in N_i$. Equipped with the above notation, we are now ready to formulate the CPF problem.

Definition 2.3 (CPF). *Consider a set of vehicles V_i; $i \in \mathcal{N}$, with dynamics* (2.1)*, together with a corresponding set of paths $y_{d_i}(\gamma_i)$ parameterized by γ_i and a formation speed assignment $v_L(\gamma_i, t)$. Assume that for each vehicle there is a feedback control law $u_i(x_i, y_{d_i}, v_L)$ such that the closed-loop systems* (2.5) *satisfy Assumption* 2.2*. Further assume that γ_i and γ_j, $j \in N_i$, are available to vehicle $i \in \mathcal{N}$. Given* $\epsilon > 0$ *arbitrarily small, derive a control law for v_{r_i} such that the PF errors $\|e_i\|$, the coordination errors $\gamma_i - \gamma_j$, and the formation speed tracking errors $\dot{\gamma}_i - v_L$ $\forall i, j \in \mathcal{N}$, converge to a ball of radius ϵ around zero as $t \to \infty$.*

3. Preliminaries and basic results.

With the setup adopted, graph theory becomes the tool par excellence for modeling the constraints imposed by the communication topology among the vehicles, as embodied in the definition of sets N_i, $i \in \mathcal{N}$. We now recall some key concepts from algebraic graph theory [24] and agreement algorithms and derive some basic tools that will be used in what follows.

3.1. Graph theory.

Let $\mathcal{G}(\mathcal{V}, \mathcal{E})$ (abbreviated $\mathcal{G}$) be the undirected graph induced by the intervehicle communication network, with $\mathcal{V}$ denoting the set of n nodes (each corresponding to a vehicle) and $\mathcal{E}$ the set of edges (each representing a data link). Nodes i and j are said to be adjacent if there is an edge between them. A path of length r between node i and node j consists of $r+1$ consecutive adjacent nodes. We say that $\mathcal{G}$ is connected when there exists a path connecting every two nodes in the graph. The adjacency matrix of a graph, denoted A, is a square matrix with rows and columns indexed by the nodes such that the i,j-entry of A is 1 if $j \in N_i$ and zero otherwise. The degree matrix D of a graph $\mathcal{G}$ is a diagonal matrix where the i,i-entry equals $|N_i|$, the cardinality of N_i. The Laplacian of a graph is defined as $L := D - A$. It is well known that L is symmetric and $L\mathbf{1} = \mathbf{0}$, where $\mathbf{1} := [1]_{n\times 1}$ and $\mathbf{0} := [0]_{n\times 1}$. If $\mathcal{G}$ is connected, then L has a simple eigenvalue at zero with an associated eigenvector $\mathbf{1}$, and the remaining eigenvalues are all positive.

We will be dealing with situations where the communication links are time-varying in the sense that links can appear and disappear (switch) due to intermittent failures and/or communication links scheduling. The mathematical setup required is described next.

A *complete graph* is a graph with an edge between each pair of nodes. A complete graph with n nodes has $\bar{n} = n(n-1)/2$ edges. Let $\mathcal{G}$ be a complete graph with edges numbered $1, \dots, \bar{n}$. Consider a communication network among n agents. To model the underlying switching communication graph, let $p = [p_i]_{\bar{n}\times 1}$, where each $p_i(t) : [0, \infty) \to \{0, 1\}$ is a piecewise-continuous time-varying binary function which indicates the existence of edge i in the graph $\mathcal{G}$ at time t. Therefore, given a switching signal $p(t)$, the dynamic communication graph $\mathcal{G}_{p(t)}$ is the pair $(\mathcal{V}, \mathcal{E}_{p(t)})$, where $p_i(t) = 1$ if $i \in \mathcal{E}_{p(t)}$ and $p_i(t) = 0$ otherwise. For example, $p(t) = [1, 0, \dots, 0]^T$ means that at time t only link number 1 is active. Denote by L_p the explicit dependence of the graph Laplacian on p and likewise for the degree matrix D_p and the adjacency matrix A_p. Further let $N_{i,p(t)}$ denote the set of the neighbors of agent i at time t.

We discard infinitely fast switchings. Formally, we let S_{dwell} denote the class of piecewise constant switching signals such that any consecutive discontinuities are separated by no less than some fixed positive constant time τ_D, the dwell time. We assume that $p(t) \in S_{dwell}$.

3.2. Brief connectivity losses (BCL). Consider the situation where the communication network changes in time so as to make the underlying dynamic communication graph $\mathcal{G}_{p(t)}$ alternately connected and disconnected. To study the impact of temporary connectivity losses on the performance of the coordination algorithms developed, we explore the concept of "brief instabilities" developed in [25]. In particular, this concept will be instrumental in capturing the percentage of time that the communication graph is not connected.

Recall that the binary value of the element p_i in p declares the existence of edge i in graph $\mathcal{G}_p$. We can thus build $2^{\bar{n}}$ graphs indexed by the different possible occurrence of vector p. Let P denote the set of all possible vectors p, and let P_c and P_{dc} denote the partitions of P that give rise to connected graphs and disconnected graphs, respectively. That is, if $p \in P_c$, then $\mathcal{G}_p$ is connected, or otherwise disconnected. Define the characteristic function of the switching signal p as

$$\chi(p) := \begin{cases} 0, & p \in P_c, \\ 1, & p \in P_{dc}. \end{cases} \tag{3.1}$$

For a given time-varying $p(t) \in S_{dwell}$, the connectivity loss time $T_p(\tau, t)$ over $[\tau, t]$ is defined as

$$T_p(t, \tau) := \int_\tau^t \chi(p(s))ds. \tag{3.2}$$

Definition 3.1 (BCL). *The communication network is said to have* BCL *if*

$$T_p(t, \tau) \le \alpha(t - \tau) + (1 - \alpha)T_0 \quad \forall t \ge \tau \ge 0 \tag{3.3}$$

for some $T_0 > 0$ and $0 \le \alpha \le 1$. In this case, $p(t) \in \mathcal{P}_{BCL}(\alpha, T_0) \subset S_{dwell}$, where $\mathcal{P}_{BCL}(\alpha, T_0)$ is identified with the set of time-varying graphs for which the connectivity loss time $T_p(\tau, t)$ satisfies (3.3).

In (3.3), α provides an asymptotic upper bound on the ratio $T_p(\tau, t)/(t - \tau)$ as $t - \tau \to \infty$ and is therefore called the *asymptotic connectivity loss rate.* When $p \in P_{dc}$ over an interval $[\tau, t]$, we have $T_p(\tau, t) = t - \tau$, and the above inequality requires that $t - \tau \le T_0$. This justifies calling T_0 the *connectivity loss upper bound.* Notice that $\alpha = 1$ means that the communications graph is never connected.

We now introduce a special coordination error vector and some preliminary results that will play an important role in the sections that follow. As will be shown later, the error state thus introduced will be zero iff the coordination states are equal. To this effect, start by stacking the coordination states in a vector $\gamma := [\gamma_i]_{n \times 1}$. Given a diagonal matrix $K > 0$, define $\beta := K^{-1}\mathbf{1}$ and the error vector

$$\tilde{\gamma} := \mathcal{L}_\beta \gamma, \tag{3.4}$$

where

$$\mathcal{L}_\beta := I - \frac{1}{\beta^T \mathbf{1}} \mathbf{1} \beta^T \tag{3.5}$$

and I is an identity matrix. The following lemma holds true.

Lemma 3.2. *The error vector $\tilde{\gamma}$, the matrix $\mathcal{L}_\beta$, and the graph Laplacian L_p satisfy the following properties:*

1. *$\mathcal{L}_\beta$ has $n - 1$ eigenvalues at 1 and a single eigenvalue at 0 with right and left eigenvectors $\mathbf{1}$ and β, respectively, such that $\mathcal{L}_\beta \mathbf{1} = \mathbf{0}$ and $\beta^T \mathcal{L}_\beta = \mathbf{0}^T$.*
2. *$\mathcal{L}_\beta K L_p = K L_p$ $\forall p \in P_c \cup P_{dc}$.*

3. $\nu\mathcal{L}_\beta^T K^{-1}\mathcal{L}_\beta\nu \leq \nu K^{-1}\nu \ \forall \nu \in \mathbb{R}^n$.
4. $\tilde{\gamma} = \mathbf{0} \Leftrightarrow \gamma \in \operatorname{span}\{\mathbf{1}\}$.
5. $\beta^T\tilde{\gamma} = 0$.
6. $L_p\tilde{\gamma} = L_p\gamma \ \forall p \in P_c \cup P_{dc}$.
7. *If* $\|\tilde{\gamma}\| < \epsilon$, *then* $|\gamma_i - \gamma_j| < \sqrt{2}\epsilon$ *and* $\|KL_p\gamma\| < n\epsilon\|K\|$.
8. *Let*

$$\lambda_{2,m} := \min_{\substack{p\in P_c\\ \mathbf{1}^T\nu=0\\ \nu^T\nu\neq 0}} \frac{\nu^T L_p \nu}{\nu^T\nu}, \qquad \lambda_m := \min_{\substack{p\in P_c\\ \beta^T\nu=0\\ \nu^T\nu\neq 0}} \frac{\nu^T L_p \nu}{\nu^T\nu}, \qquad \bar{\lambda}_m := \min_{\substack{p\in P_c\cup P_{dc}\\ L_p\nu\neq 0\\ \nu^T\nu\neq 0}} \frac{\nu^T L_p \nu}{\nu^T\nu}.$$

Then, $\lambda_m = \frac{(\beta^T\mathbf{1})^2}{n\beta^T\beta}\lambda_{2,m} > 0$ *and* $\bar{\lambda}_m > 0$.

9. *If* $z = L_{p(t)}\gamma$, *then the ith component of* z *is* $z_i = \sum_{j\in N_{i,p(t)}} \gamma_i - \gamma_j$.
10. $\|\mathcal{L}_\beta \mathbf{v}_L(\gamma, t)\| \leq \sqrt{n}\min(2v_M, \sqrt{2}\,l\|\tilde{\gamma}\|)$, *where* $\mathbf{v}_L(\gamma, t) = [v_L(\gamma_i, t)]_{n\times 1}$.

Proof. See the appendix. □

Property 4 allows for the conclusion that if $\tilde{\gamma}$ tends to zero, then $|\gamma_i - \gamma_j| \to 0$ $\forall i, j \in \mathcal{N}$, as $t \to \infty$ and coordination is achieved. Property 7 gives a bound on the coordination errors $\gamma_i - \gamma_j$ given a bound on the error vector $\tilde{\gamma}$. In the literature, the connectivity of a graph with Laplacian L is defined as the second smallest eigenvalue λ_2 of L. The term $\lambda_{2,m}$ defined in property 8 is an extension of the concept of connectivity in a collective sense, defined as the smallest graph connectivity over all connected graphs $\mathcal{G}_p$. Given λ_m, the lower bound estimate $\tilde{\gamma}^T L_p\tilde{\gamma} \geq \lambda_m\tilde{\gamma}^T\tilde{\gamma}$, when $p \in P_c$, applies. An identical interpretation applies to $\bar{\lambda}_m$. Notice from property 9 that if the control signal of vehicle i is computed as a function of z_i, then the proposed control law meets the communication constraints embodied in the sets N_i.

3.3. Uniformly connected in mean topology. In the previous situation, we considered the case where the communication graph changes in time, alternating between connected and disconnected graphs. We now address a more general case where the communication graph may even fail to be connected at any instant of time; however, we assume there is a finite time $T > 0$ such that over any interval of length T the union of the different graphs is somehow connected. This statement is made precise in what follows. We now present some key results for the time-varying communication graph that borrow from [37], [38], [40].

Let $\mathcal{G}_i$, $i = 1, \ldots, q$, be q graphs defined on n nodes and denote by L_i their corresponding graph Laplacians. Define the *union graph* $\mathcal{G} = \cup_i \mathcal{G}_i$ as the graph whose edges are obtained from the union of the edges $\mathcal{E}_i$ of $\mathcal{G}_i$, $i = 1, \ldots, q$. If $\mathcal{G}$ is connected, $L = \sum_i L_i$ has a single eigenvalue at 0 with eigenvector $\mathbf{1}$. Notice that L is not necessarily the Laplacian of $\mathcal{G}$, because for an edge e, if $e \in \mathcal{E}_i$ and $e \in \mathcal{E}_j$ for $i \neq j$, then e is counted twice in L through $L_i + L_j$, while we consider only one link in $\mathcal{G}$ as representative of e. However, L has the same rank properties as the Laplacian of $\mathcal{G}$. Since $p \in S_{dwell}$ (only a finite number of switchings are allowed over any bounded time interval), the union graph is defined over time intervals in the obvious manner. Formally, given two real numbers $0 \leq t_1 \leq t_2$, the union graph $\mathcal{G}([t_1, t_2))$ is the graph whose edges are obtained from the union of the edges $\mathcal{E}_{p(t)}$ of graph $\mathcal{G}_{p(t)}$ for $t \in [t_1, t_2)$.

Definition 3.3 (uniformly connected in mean (UCM)). *A switching communication graph* $\mathcal{G}_{p(t)}$ *is* UCM *if there exists* $T > 0$ *such that for every* $t \geq 0$ *the union graph* $\mathcal{G}([t, t + T))$ *is connected.*

For a given $t > 0$, let $s_0 = t$ and the sequence s_i, $i = 1, \ldots, q$, be the time instants at which switching happens over the interval $[t, t+T)$. If the switching communication graph is UCM, then the union graph $\cup_{i=0}^{q} \mathcal{G}_i$ is connected and $\sum_{i=0}^{q} L_{p(s_i)}$ has a single eigenvalue at origin with eigenvector $\mathbf{1}$.

Consider the linear time-varying system

$$\dot{\gamma} = -KL_p\gamma, \tag{3.6}$$

where K is a positive definite diagonal matrix and L_p is the Laplacian matrix of a dynamic graph $\mathcal{G}_p$. The following theorem holds; see, for example, [38].

Theorem 3.4 (agreement). *Coordination (agreement) among the variables γ_i with dynamics* (3.6) *is achieved uniformly exponentially if the switching communication graph $\mathcal{G}_{p(t)}$ is UCM. That is, under this connectivity condition, all the coordination errors $\gamma_{ij}(t)$ converge to zero and $\dot{\gamma}_i \to 0$ as $t \to \infty$.*

We now consider the delayed version of (3.6). Let the coordination states γ_i evolve according to

$$\dot{\gamma}(t) = -KD_{p(t)}\gamma(t) + KA_{p(t)}\gamma(t - \tau), \tag{3.7}$$

where $D_{p(t)}$ and $A_{p(t)}$ are the degree matrix and the adjacency matrix of $\mathcal{G}_{p(t)}$, respectively. The following theorem can be derived from the results in [40].

Theorem 3.5 (agreement-delayed information). *The variables γ_i with dynamics* (3.7) *agree uniformly exponentially for $\tau \geq 0$ if the switching communication graph $\mathcal{G}_{p(t)}$ is UCM, that is, under this connectivity condition, all the coordination states $\gamma_i(t)$ converge to the same value and $\dot{\gamma}_i \to 0$ as $t \to \infty$.*

A version of Definition 3.3 for directed graphs was first introduced in [38], where the term "uniformly quasi-strongly connected" was used. Here, we adapt this definition to undirected graphs, thus the term "uniformly connected in mean" seems to be more adequate. It is interesting to point out that Theorem 3.4 follows naturally from the work in [38] or from Theorem 3.4 in [37], which recovers some of the results in [38] for linear systems. Theorem 3.4 can also be derived from Theorem 1 in [40] by using the fact that $p(t) \in S_{dwell}$ with a dwell time $\tau_D > 0$. Finally, Theorem 3.5 can be derived from Theorem 2 in [40] by noticing that $-KL_p$ is a matrix with nonnegative off-diagonal elements (Metzler matrix) with all its row-sums equal to zero.

4. System interconnections. Systems with brief instabilities. This section introduces a lemma that will be instrumental in deriving the performance measure (error decay rate) associated with the coordination algorithm that will be later derived for multivehicle systems communicating over networks with BCLs (Definition 3.1). Here, we avail ourselves of some important results on brief instabilities [25]. We start with basic definitions. A switching linear system $S : \dot{x} = A_p x + B_p u$ is a dynamical system, where A_p and B_p are functions of some time-varying vector function $p(t)$. The characteristic function of S, denoted χ, is defined as $\chi(p) = 0$ if S is stable and as $\chi(p) = 1$ otherwise. Let the instability time $T_p(t, \tau)$ of S be defined in a manner similar to (3.2). Then, S is said to have *brief instabilities* with instability bound T_0 and asymptotic instability rate α if T_p satisfies (3.3).

Lemma 4.1 (system interconnection and brief instabilities). *Consider the coupled system consisting of two subsystems*

$$\begin{aligned}\dot{z}_1 &= \phi_1(t, z_1, z_2, u_1),\\ \dot{z}_2 &= \phi_2(t, z_1, z_2, u_2),\end{aligned}$$

denoted System 1 and System 2, respectively, where z_1 and z_2 denote the state vectors and u_1 and u_2 the inputs. Assume there exist Lyapunov functions $V_1(t, z_1)$ and $V_2(t, z_2)$ satisfying

$$\begin{aligned} \underline{\alpha}_1\|z_1\|^2 &\le V_1 \le \bar{\alpha}_1\|z_1\|^2, \\ \underline{\alpha}_2\|z_2\|^2 &\le V_2 \le \bar{\alpha}_2\|z_2\|^2 \end{aligned} \tag{4.1}$$

and

$$\begin{aligned} \frac{\partial V_1}{\partial t} + \frac{\partial V_1}{\partial z_1}\phi_1 &\le -\lambda_1 V_1 + \rho_1\|z_2\|^2 + u_1^2, \\ \frac{\partial V_2}{\partial t} + \frac{\partial V_2}{\partial z_2}\phi_2 &\le -\lambda_2(t) V_2 + \rho_2\|z_1\|^2 + u_2^2, \end{aligned} \tag{4.2}$$

where $\underline{\alpha}_i$, $\bar{\alpha}_i$, ρ_i, $i = 1, 2$, and λ_1 are positive values. Let system 2 have brief instabilities characterized by

$$\chi(p(t)) = \begin{cases} 0, & \lambda_2(p(t)) = \lambda_2, \\ 1, & \lambda_2(p(t)) = -\tilde{\lambda}_2, \end{cases} \tag{4.3}$$

where $\lambda_2 > 0$, $\tilde{\lambda}_2 \ge 0$, with asymptotic instability rate α and instability bound T_0. Define

$$\lambda_0 := \frac{1}{2}(\lambda_1 + \lambda_2) - \sqrt{\frac{1}{4}(\lambda_1 + \lambda_2)^2 - \lambda_1\lambda_2 + \frac{\rho_1\rho_2}{\underline{\alpha}_1\underline{\alpha}_2}} \tag{4.4}$$

that satisfies

$$\min(\lambda_1, \lambda_2) - \sqrt{\frac{\rho_1\rho_2}{\underline{\alpha}_1\underline{\alpha}_2}} \le \lambda_0 \le \max(\lambda_1, \lambda_2) - \sqrt{\frac{\rho_1\rho_2}{\underline{\alpha}_1\underline{\alpha}_2}}.$$

Assume that $\alpha < \lambda_0/(\lambda_2 + \tilde{\lambda}_2)$ and

$$\rho_1\rho_2 < \underline{\alpha}_1\underline{\alpha}_2\lambda_1\lambda_2. \tag{4.5}$$

Then,

1. *the interconnected system is ISS with respect to state $z = \mathrm{col}(z_1, z_2)$ and input $u = \mathrm{col}(u_1, u_2)$.*
2. *there is a Lyapunov function $V(t, z)$ such that*

$$\begin{aligned} \underline{\alpha}\|z\|^2 &\le V \le \bar{\alpha}\|z\|^2, \\ V(t) &\le cV(t_0)e^{-\lambda(t-t_0)} + g\sup_{[t_0,t]} u^2, \end{aligned} \tag{4.6}$$

where $c = e^{(\lambda_2+\tilde{\lambda}_2)(1-\alpha)T_0}$, $g = \frac{c}{\lambda}\max(1, \underline{\alpha}_1(\lambda_1 - \lambda_0)/\rho_2)$, and the rate of convergence λ is given by $\lambda = \lambda_0 - \alpha(\lambda_2 + \tilde{\lambda}_2)$.

In particular, if $\rho_2 = 0$ and $\rho_1 > 0$, then the interconnected system takes a cascade form and is ISS with input u and state z. Furthermore, the system exhibits convergence rate $\lambda = \min(\lambda_1, (1 - \alpha)\lambda_2 - \alpha\tilde{\lambda}_2)$. The conclusions are also valid with $\alpha = 0$ for the case where system 2 has no instabilities, that is, $\lambda_2(t) = \lambda_2$.

Proof. An indication of the proof for the case where ρ_1 and ρ_2 are nonzero is given next. See the appendix for the proof in the general case.

Define $V = V_1 + aV_2$ for some $a > 0$ to be chosen later. Taking the derivative of V yields

$$\dot{V} \leq -\left(\lambda_1 - \frac{a\rho_2}{\underline{\alpha}_1}\right) V_1 - a\left(\lambda_2(t) - \frac{\rho_1}{a\underline{\alpha}_2}\right) V_2 + g\|d\|^2,$$

where $g = \max(1, a)$. Given any constant $\lambda_2 > 0$, there exists $a > 0$ such that

$$\lambda_1 - \frac{a\rho_2}{\underline{\alpha}_1} = \lambda_2 - \frac{\rho_1}{a\underline{\alpha}_2} \tag{4.7}$$

if (4.5) is satisfied (small-gain condition). Then $\dot{V} \leq -\lambda_0 V + g\|d\|^2$, where λ_0 is given by (4.4), and the interconnected system is ISS with input d. Furthermore, its convergence rate is $\lambda = \lambda_0$.

Consider now the situation where $\lambda_2(t)$ is time-varying. In this case, $\dot{V} \leq -\lambda_0 V + a(\lambda_2 - \lambda_2(t))V_2 + g\|d\|$. Because system 2 has brief instabilities with characteristic function $\chi(p)$, using the relationship $aV_2 = V - V_1$ yields

$$\dot{V} \leq -(\lambda_0 - \lambda_3\chi(p(t)))V + g\|d\|^2,$$

where $\lambda_3 := \lambda_2 + \tilde{\lambda}_2$. Integrating the above differential inequalities, it can be shown that

$$V(t) \leq V(t_0)e^{-\lambda_0(t-t_0)+\lambda_3 T_p} + g \sup\nolimits_{[t_0,t]} \|d\|^2 \int_{t_0}^{t} e^{-\lambda_0(t-\tau)+\lambda_3 T_p} d\tau.$$

Using (3.3) concludes the proof. □

It is interesting to notice how the lemma invokes two conditions: (i) the small-gain condition (4.5), which is sufficient to guarantee that the results stated hold true when system 2 is stable, and (ii) the extra inequality $\alpha < \lambda_0/(\lambda_2 + \tilde{\lambda}_2)$, that must also be satisfied when system 2 has brief instabilities. In this respect, the above lemma generalizes the results derived in [27] for the case where system 2 has no brief instabilities. As an example of application of the lemma, assume $\lambda_1 = \lambda_2 = \tilde{\lambda}_2 = \frac{1}{k}\sqrt{\frac{\rho_1\rho_2}{\underline{\alpha}_1\underline{\alpha}_2}}$, where $0 < k < 1$. Then the small-gain condition is satisfied and the interconnected system of the lemma above is ISS if $\alpha < \frac{1-k}{2}$, which is smaller than 0.5 for any admissible k.

Equipped with the results derived so far, the next two sections offer solutions to the CPF problem formulated in section 2.

5. CPF in the absence of communication delays. Consider the coordination control problem introduced in section 2 with a switching communication topology parameterized by $p : [0, \infty) \to \{0, 1\}$ and with no communication delays. Recall that the coordination states γ_i are governed by (2.8). Inspired by the work in [28], [61], we propose the following decentralized feedback law for the reference speeds v_{r_i} as a function of the information obtained from the neighboring vehicles:

$$v_{r_i} = v_L - k_i \sum_{j \in N_{i,p(t)}} (\gamma_i(t) - \gamma_j(t)), \tag{5.1}$$

where $v_L(\gamma_i, t)$ is the common, nominal speed assigned to the fleet of vehicles and $k_i > 0$. Let $k_m := \min_i k_i$ and $k_M := \max_i k_i$. Notice that with this choice of control law, the term $\tilde{v}_{r_i} = v_{r_i} - v_L$, for which the time derivative is not available, is given by

$$\tilde{v}_{r_i} = -k_i \sum_{j \in N_{i,p(t)}} (\gamma_i(t) - \gamma_j(t)). \tag{5.2}$$

Using (2.8), (5.1), and property 9 of Lemma 3.2, the coordination control closed-loop system can be written in vector form as

$$\dot{\gamma} = -KL_{p(t)}\gamma + \mathbf{v}_L(\gamma, t) + g_\eta \eta, \tag{5.3}$$

where $K = \text{diag}[k_i]$. The auxiliary term g_η was added for simplicity of exposition: $g_\eta = 1$ when the closed-loop PF system is of Type I (η is considered a state), and $g_\eta = 0$ when the PF system is of Type II ($\eta = \mathbf{0}$); see Assumption 2.2. Using properties 2 and 6 of Lemma 3.2, the coordination dynamics (5.3) take the form

$$\dot{\tilde{\gamma}} = -KL_p\tilde{\gamma} + \mathcal{L}_\beta \mathbf{v}_L(\gamma, t) + g_\eta \mathcal{L}_\beta \eta. \tag{5.4}$$

Notice from (5.3) that η can be viewed as a coupling term from the PF to the coordination dynamics.

At this stage, in preparation for the following sections, we state a lemma on an ISS property that applies to a collection of PF systems.

Lemma 5.1. *Consider n PF subsystems, each satisfying Assumption 2.2, and let $\zeta = [\zeta_i]_{n\times 1}$. Then there exists a single Lyapunov function V_1 satisfying*

$$\begin{aligned} &\underline{\alpha}_1\|\zeta\|^2 \le V_1 \le \bar{\alpha}_1\|\zeta\|^2, \\ &\dot{V}_1 \le -\lambda_1 V_1 + \rho_1 n^2 k_M^2 \|\tilde{\gamma}\|^2 + u_1^2, \end{aligned} \tag{5.5}$$

where $u_1^2 := \sum_{i=1}^n d_i^2$. In addition, the ISS property

$$\|\eta(t)\| \le \|\zeta(t)\| \le e^{-\bar{\lambda}_1(t-t_0)}\|\zeta(t_0)\| + \bar{\rho}_1 \sup_{\tau\in[t_0,t)} \|\tilde{\gamma}\| + \bar{\rho}_2\|u_1\| \tag{5.6}$$

holds with $\bar{\lambda}_1 = \frac{\underline{\alpha}_1}{2\bar{\alpha}_1}\lambda_1$, $\bar{\rho}_1 = \sqrt{\frac{\rho_1 n^2 k_M^2}{\lambda_1 \underline{\alpha}_1}}$, and $\bar{\rho}_2 = \frac{1}{\sqrt{\lambda_1 \underline{\alpha}_1}}$.

Proof. See the appendix. □

Close inspection of the ISS property (5.6) and the dynamics (5.4) shows that the PF and coordination systems form a feedback interconnected system.

To deal with switching communication topologies, two approaches are introduced next: "uniform switching topologies" and "brief connectivity losses," as defined in section 2. We now derive conditions under which the overall closed-loop system consisting of the PF and coordination subsystems is stable. We also derive some convergence properties for the complete system.

5.1. UCM topology.

This section addresses the case where the communication network changes but the underlying communication graph is UCM (see Definition 3.3). Recall in this case that there is $T > 0$ such that for any $t \ge 0$, the union graph $\mathcal{G}([t, t+T))$ is connected. The section starts with some preliminary results leading to the statement of Theorem 5.2, a proof of which is included in the appendix.

Consider the unforced coordination closed-loop dynamics derived from (5.4), that is,

$$\dot{\tilde{\gamma}} = -KL_p\tilde{\gamma}. \tag{5.7}$$

First, we will show that if the switching communication graph is UCM (with parameter $T > 0$), then $\forall t > 0$, $\exists \tau \in [t, t+T)$, such that $L_{p(\tau)}\tilde{\gamma}(\tau) \ne \mathbf{0}$. To this effect, we let $V = \frac{1}{2}\tilde{\gamma}^T K^{-1}\tilde{\gamma}$ whose time derivative along the solutions of (5.7) is

$$\dot{V} = -\tilde{\gamma}^T L_p \tilde{\gamma}.$$

Notice that $\dot{V}$ is negative semidefinite whether the graph is connected or not. Thus, $\tilde{\gamma}$ remains bounded. Consider now the sequence s_i, $i = 1, \dots, q$, of switching times in the interval $[t, t + T)$, with $t < s_1 < s_q < t + T$ and $s_i \leq s_{i+1} - \tau_D$, $i = 1, \dots, q - 1$, where τ_D is the dwell time. Let $s_0 = \min(t, s_1 - \tau_D)$ and $T_1 = \max(s_q + \tau_D, t + T) - s_0$. With this construction $T \leq T_1 \leq T + 2\tau_D$, $s_1 - \tau_D \geq s_0$, and $s_q + \tau_D \leq s_0 + T_1$. We now show that[2] $\exists \tau \in \mathbb{T} := [s_0, s_0 + T_1)$ such that $L_{p(\tau)}\tilde{\gamma}(\tau) \neq \mathbf{0}$.

Assume by contradiction that $L_p\tilde{\gamma} = \mathbf{0}$ $\forall \tau \in \mathbb{T}$ and discard the trivial solution $\tilde{\gamma} = \mathbf{0}$. Then, from (5.7) it follows that $\dot{\tilde{\gamma}} = \mathbf{0}$; that is, $\tilde{\gamma}$ remains unchanged over $\mathbb{T}$. Therefore,

$$\mathbf{0} = \sum_{i=0}^{q} L_{p(s_i)}\tilde{\gamma}(s_i) = \left(\sum_{i=0}^{q} L_{p(s_i)}\right)\tilde{\gamma}(s_0).$$

As shown in section 3, since the graph is UCM the matrix $\sum_{i=0}^{q} L_{p(s_i)}$ has rank $n - 1$ and its kernel is span$\{\mathbf{1}\}$. As a consequence, $\tilde{\gamma}(s_0) \in$ span$\{\mathbf{1}\}$, which contradicts the fact that $\beta^T\tilde{\gamma} = 0$.

Without loss of generality, assume $L_{p(s_0)}\tilde{\gamma}(s_0) \neq \mathbf{0}$ and define $\mathbb{T}_D := [s_0, s_0 + \tau_D)$. Clearly, $\forall \bar{t} \in \mathbb{T}_D$ the inequality $L_{p(\bar{t})}\tilde{\gamma}(\bar{t}) \neq \mathbf{0}$ applies because (5.7) is a linear time invariant system during the interval considered and its solutions cannot tend to zero in finite time. It follows that

$$\dot{V}(\bar{t}) \leq \begin{cases} -2k_m\bar{\lambda}_m V(\bar{t}), & \bar{t} \in \mathbb{T}_D, \\ 0, & \bar{t} \in \mathbb{T}\backslash\mathbb{T}_D, \end{cases} \tag{5.8}$$

with $\bar{\lambda}_m$ as defined in property 8 of Lemma 3.2. We can now conclude that system (5.7) with UCM switching communication graphs has brief instabilities with asymptotic instability rate $\bar{\alpha} = 1 - \tau_D/T_1 \leq 1 - \tau_D/(T + 2\tau_D)$ and instability upper bound $\bar{T}_0 = T_1 - \tau_D \leq T + \tau_D$. That is, if a characteristic function $\bar{\chi}$ is defined as

$$\bar{\chi}(t) = \begin{cases} 0, & t \in \mathbb{T}_D, \\ 1, & t \in \mathbb{T}\backslash\mathbb{T}_D, \end{cases}$$

then $\dot{V}(t) \leq -2k_m\bar{\lambda}_m(1 - \bar{\chi}(t))V(t)$. Integrating this differential inequality yields

$$V(t) \leq cV(\tau)e^{-2\lambda_\alpha(t-\tau)} \quad \forall t \geq \tau \geq 0,$$

with

$$\lambda_\alpha = (1 - \bar{\alpha})k_m\bar{\lambda}_m, \quad c = e^{2\lambda_\alpha\bar{T}_0}, \tag{5.9}$$

and where we used the fact that

$$\int_t^\tau \bar{\chi}(s)ds \leq \bar{\alpha}(t - \tau) + (1 - \bar{\alpha})\bar{T}_0 \quad \forall t \geq \tau \geq 0.$$

Therefore, $\|\tilde{\gamma}(t)\| \leq c_1e^{-\lambda_\alpha(t-\tau)}\|\tilde{\gamma}(\tau)\|$ and

$$\|\Phi_p(t, \tau)\| \leq c_1e^{-\lambda_\alpha(t-\tau)}, \tag{5.10}$$

where $\Phi_p(t, \tau)$ denotes the state transition matrix of (5.7) and $c_1 = \sqrt{\frac{ck_M}{k_m}}$. Notice that the above inequality is valid for all $p(t) \in S_{dwell}$ such that the graph $\mathcal{G}_p$ is UCM.

[2] Notice that if $\exists \tau \in \mathbb{T}$ such that $L_{p(\tau)}\tilde{\gamma}(\tau) \neq \mathbf{0}$, then $\exists \tau_1 \in [t, t + T)$ such that $L_{p(\tau_1)}\tilde{\gamma}(\tau_1) \neq \mathbf{0}$ because $t \leq s_0 + \tau_D$ and $t + T \geq s_0 + T - \tau_D$.

For a given switching signal $p(t)$, input $\eta(t)$, and initial state $\gamma(t_0)$, the solution of (5.4) is given by (see [53, p. 87])

$$\tilde{\gamma}(t) = \Phi_p(t, t_0)\tilde{\gamma}(t_0) + \int_{t_0}^{t} \Phi_p(t, \tau)\mathcal{L}_\beta \mathbf{v}_L(\gamma(\tau), \tau)d\tau + g_\eta \int_{t_0}^{t} \Phi_p(t, \tau)\mathcal{L}_\beta \eta(\tau)d\tau \quad \forall t \geq t_0.$$

Letting

$$\bar{\lambda}_\alpha = \lambda_\alpha - c_1\sqrt{2n}l = \lambda_\alpha - e^{\lambda_\alpha \bar{T}_0} l \sqrt{2n\frac{k_M}{k_m}} \tag{5.11}$$

and using (5.10) and property 10 of Lemma 3.2, an upper bound for $\tilde{\gamma}(t)$ can be derived as

$$\|\tilde{\gamma}(t)\| \leq c_1 e^{-\lambda_\alpha(t-t_0)}\|\tilde{\gamma}(t_0)\| + c_1 l\sqrt{2n}\int_{t_0}^{t} e^{-\lambda_\alpha(t-\tau)}\|\tilde{\gamma}(\tau)\|d\tau + g_\eta \frac{c_1}{\lambda_\alpha} \sup_{\tau\in[t_0,t)} \|\eta(\tau)\| \tag{5.12}$$

if $\bar{\lambda}_\alpha > 0$, and

$$\|\tilde{\gamma}(t)\| \leq c_1 e^{-\lambda_\alpha(t-t_0)}\|\tilde{\gamma}(t_0)\| + \frac{2v_M c_1\sqrt{n}}{\lambda_\alpha} + g_\eta \frac{c_1}{\lambda_\alpha} \sup_{\tau\in[t_0,t)} \|\eta(\tau)\| \tag{5.13}$$

otherwise. It is now straightforward to multiply both sides of (5.12) by $e^{\lambda_\alpha t}$ and to use the Gronwall–Bellman theorem [31, p. 66] to arrive at

$$\|\tilde{\gamma}(t)\| \leq c_1 e^{-\bar{\lambda}_\alpha(t-t_0)}\|\tilde{\gamma}(t_0)\| + g_\eta \frac{c_1}{\bar{\lambda}_\alpha} \sup \|\eta(\tau)\| \tag{5.14}$$

provided that $\bar{\lambda}_\alpha > 0$. Notice from (5.11) that $\bar{\lambda}_\alpha$ cannot be made arbitrarily large. It can be shown that there are control gains ($k_m = k_M$) that make $\bar{\lambda}_\alpha > 0$ if the Lipschitz constant l of v_L satisfies

$$l < \frac{1}{(T + \tau_D)\sqrt{2ne}}. \tag{5.15}$$

For each such l, the corresponding maximum value of $\bar{\lambda}_\alpha$ can be easily computed.

Equipped with these introductory results, we now state the main theorem of this section.

THEOREM 5.2 (CPF with UCM). *Consider the interconnected system* Σ *depicted in Figure* 3, *consisting of* n *PF subsystems satisfying Assumption* 2.2 *together with the coordination control (CC) subsystem* (5.3) *supported by a communication network*

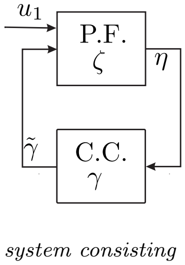


FIG. 3. Σ: *Overall closed-loop system consisting of the PF and CC subsystems.*

that is UCM with parameter T and switching dwell time τ_D. Then, Σ is input-to-state practical stable (ISpS) with respect to the states $\tilde{\gamma}$ and ζ, the input u_1, and the constant $2v_M c_1\sqrt{n}/\lambda_\alpha$ if

$$\begin{cases} c_1\sqrt{\frac{\rho_1 n^2 k_M^2}{\lambda_1\bar{\alpha}_1}} < \bar{\lambda}_0, & \textit{PF of Type } \mathrm{I}, \\ \textit{always}, & \textit{PF of Type } \mathrm{II}, \end{cases} \tag{5.16}$$

where $\bar{\lambda}_0 = \lambda_\alpha$ as defined in (5.9). *If* (5.15) *holds, the control gains can be chosen such that $\bar{\lambda}_\alpha > 0$. In this case, Σ is ISS with respect to the states $\tilde{\gamma}$ and ζ and input u_1 under condition* (5.16) *with $\bar{\lambda}_0 = \bar{\lambda}_\alpha$ as defined in* (5.11). *Furthermore, the PF error vectors e_i, the speed tracking errors $|\dot{\gamma}_i - v_L|$, and the coordination errors $|\gamma_i - \gamma_j|$ $\forall i, j \in \mathcal{N}$ converge exponentially fast to some ball around zero as $t \to \infty$, with rate at least* $\min(\bar{\lambda}_0, \bar{\lambda}_1)$.

Proof. A proof of (5.15) is given in the appendix. Using the ISS version of the small-gain theorem for the interconnection of (5.14) and (5.6) in the case of $\bar{\lambda}_\alpha > 0$, and for the interconnection of (5.13) and (5.6) otherwise, leads to the result. □

From the above, under the UCM assumption, it follows that the complete CPF control system is ISS if condition (5.15) is satisfied. In the absence of disturbances and noise, the origin of the system becomes globally asymptotically stable (in fact, exponentially stable). In the case when condition (5.15) is not satisfied, all that can be shown is that the complete system is ISpS.

5.2. BCLs. This section addresses the situation where the communication network has BCLs; see Definition 3.1. In this case the underlying communication graph switches between connected and disconnected configurations with known asymptotic connectivity loss rate α and connectivity loss upper bound T_o.

The following result provides conditions under which the overall closed-loop system consisting of the PF and coordination subsystems is ISS.

THEOREM 5.3 (CPF with BCLs). *Consider the interconnected system Σ depicted in Figure* 3, *consisting of n PF subsystems that satisfy Assumption* 2.2 *and the coordination subsystem* (5.3) *with a communication network subjected to BCLs characterized by* (3.3). *Let $\lambda_2 := k_m\lambda_m - \frac{k_M l\sqrt{2n}}{k_m}$. Define $k_g := \frac{k_m\lambda_2^2}{n^2 k_M^3}$ and*

$$\lambda_0 = \tilde{\lambda}_0 - \sqrt{\tilde{\lambda}_0^2 - \lambda_1\lambda_2\left(1 - \frac{\rho_1}{k_g\underline{\alpha}_1\lambda_1}\right)},$$

where $\tilde{\lambda}_0 = \frac{1}{2}(\lambda_1 + \lambda_2)$ and λ_m is defined in Lemma 3.2, *property* 8. *Assume*

$$\frac{k_m^2}{k_M} > \frac{l\sqrt{2n}}{\lambda_m}. \tag{5.17}$$

Further assume the following conditions hold:

(a) [PF of Type I] *The asymptotic connectivity losses rate α satisfies*

$$\alpha < \frac{\lambda_0}{2k_m\lambda_m}$$

and

$$\frac{\rho_1}{\underline{\alpha}_1\lambda_1} < k_g.$$

(b) [PF of Type II] $\alpha < 1 - \frac{k_M l\sqrt{2n}}{\lambda_m k_m^2}$.

Then, Σ is ISS with respect to the states $\tilde{\gamma}$ and ζ and input u_1 (see Figure 3). Furthermore, the PF error vectors e_i, the speed tracking errors $|\dot{\gamma}_i - v_L|$, and the coordination errors $|\gamma_i - \gamma_j|$ $\forall i, j \in \mathcal{N}$ converge exponentially fast to some ball around zero (depending on the size of u_1) as $t \to \infty$, with rate at least

$$\lambda = \begin{cases} \lambda_0 - 2\alpha k_m \lambda_m, & \textit{PF of Type } \mathrm{I}, \\ \min(\lambda_1, \lambda_2 - 2\alpha k_m \lambda_m), & \textit{PF of Type } \mathrm{II}. \end{cases}$$

Proof. Choose the Lyapunov candidate function

$$V_2 := \frac{1}{2}\tilde{\gamma}^T K^{-1}\tilde{\gamma}$$

whose time derivative along the solutions of (5.4) is

$$\begin{aligned} \dot{V}_2 &= -\tilde{\gamma}^T L_p \tilde{\gamma} + \tilde{\gamma}^T K^{-1}\mathcal{L}_\beta \mathbf{v}_L(\gamma, t) + g_\eta \tilde{\gamma}^T K^{-1}\mathcal{L}_\beta \eta \\ &\leq -\tilde{\gamma}^T L_p \tilde{\gamma} + \frac{l\sqrt{2n}}{k_m}\|\tilde{\gamma}\|^2 + g_\eta \theta_1 \tilde{\gamma}^T K^{-1}\tilde{\gamma} + \frac{g_\eta}{4\theta_1}\eta^T \mathcal{L}_\beta^T K^{-1}\mathcal{L}_\beta \eta, \end{aligned}$$

where we used Young's inequality and property 10 of Lemma 3.2. Using properties 3 and 8 of Lemma 3.2, the above inequality yields

$$\dot{V}_2 \leq \begin{cases} -\lambda_2 V_2 + \rho_2\|\eta\|^2, & p \in P_c, \\ \tilde{\lambda}_2 V_2 + \rho_2\|\eta\|^2, & p \in P_{dc}, \end{cases} \tag{5.18}$$

with $\tilde{\lambda}_2 = \frac{2k_M l\sqrt{2n}}{k_m} + 2g_\eta\theta_1$, $\lambda_2 = 2\lambda_m k_m - \tilde{\lambda}_2$, $\rho_2 = \frac{g_\eta}{4k_m\theta_1}$. In order for λ_2 and $\tilde{\lambda}_2$ to be positive, θ_1 must satisfy $0 < \theta_1 < \lambda_m k_m - \frac{l\sqrt{2n}k_M}{k_m}$. It is straightforward to check that this condition holds if $\frac{k_m^2}{k_M} > \frac{l\sqrt{2n}}{\lambda_m}$.

Close inspection of (5.5) and (5.18) shows that the PF and coordination subsystems form a feedback interconnected system with η and $\tilde{\gamma}$ as interacting signals, as shown in Figure 3. We now use Lemma 4.1 and the fact that the coordination subsystem has BCLs as defined in (3.3) to find conditions under which the interconnected system is ISS from input u_1. We consider the cases where the PF algorithms are of Type I or II.

[*PF of Type* I] Consider the feedback interconnection of (5.5) and (5.18) for the case where $g_\eta = 1$, that is, with $\rho_2 > 0$. Resorting to Lemma 4.1 for interconnected systems with brief instabilities and applying the small-gain condition (4.5), we obtain

$$(\rho_1 n^2 k_M^2)\left(\frac{1}{4k_m\theta_1}\right) < (\underline{\alpha}_1)\left(\frac{1}{2k_M}\right)(\lambda_1)\left(2\lambda_m k_m - 2\frac{l\sqrt{2n}k_M}{k_m} - 2\theta_1\right),$$

or equivalently,

$$\frac{\rho_1}{\underline{\alpha}_1\lambda_1} < \frac{4k_m}{n^2 k_M^3}\theta_1\left(\lambda_m k_m - \frac{l\sqrt{2n}k_M}{k_m} - \theta_1\right),$$

the right-hand side of which is maximized for $\theta_1 = \frac{1}{2k_m}(\lambda_m k_m^2 - l\sqrt{2n}k_M)$. Inserting the latter value of θ_1 in the inequality above, the conditions of the theorem for PF strategies of Type I follow immediately.

[*PF of Type* II] In this case the interconnection of (5.5) and (5.18) takes a cascade form, that is, $g_\eta = 0$ and (5.18) simplifies to

$$\dot{V}_2 \leq \begin{cases} -\lambda_2 V_2, & p \in P_c, \\ \tilde{\lambda}_2 V_2, & p \in P_{dc}, \end{cases}$$

where $\tilde{\lambda}_2 = 2\frac{l\sqrt{2n}k_M}{k_m}$ and $\lambda_2 = 2\lambda_m k_m - \tilde{\lambda}_2$. Using Lemma 4.1 with $\rho_2 = 0$ the conditions of the theorem for PF strategies of Type I are easily obtained. □

At this point, it is interesting to work out a simple numerical example to illustrate some of the results derived. To this effect, consider the CPF problem for three vehicles ($n = 3$). In this case, $\lambda_m = 1$. We consider both the case where the speed profile v_L is constant and the case where $v_L(\gamma_i) = 2 + \sin(\gamma_i)$, for which $l = 0$ and $l = 1$, respectively. Choose $K = 2\sqrt{6}I_3$ to guarantee condition (5.17) for both cases of v_L. Further assume that the ISS property of the PF subsystem is satisfied with $\lambda_1 = \sqrt{6}$ and $\underline{\alpha}_1 = \sqrt{6}$. It is now straightforward to compute the following parameters consecutively. For $l = 0$: $\lambda_2 = 2\sqrt{6}$, $k_g = 4/36$, and $\tilde{\lambda}_0 = 3/2\sqrt{6}$. The small-gain condition (4.5) will require that $\rho_1 < 4/6$. For $l = 1$: $\lambda_2 = \sqrt{6}$, $k_g = 1/36$, and $\tilde{\lambda}_0 = \sqrt{6}$. The same small-gain condition will yield $\rho_1 < 1/6$ in this case. As expected, ρ_1 (which can be viewed as a stability margin) is reduced when the v_L depends on the path parameter. Let $\rho_1 = 1/24$ to ensure stability for both cases of v_L above. We can now compute $\lambda_0 = 0.54\sqrt{6}$ for $l = 0$ and $\lambda_0 = 0.5\sqrt{6}$ for $l = 1$. It follows from the above that when PF is of Type I the interconnected system will be ISS if the asymptotic connectivity loss rate is $\alpha < 13.5\%$ for $l = 0$ and $\alpha < 12.5\%$ for $l = 1$. When PF is of Type II, the bounds are relaxed to $\alpha < 100\%$ for $l = 0$ and $\alpha < 50\%$ for $l = 1$. Better convergence rates could be guaranteed if one were to aim for ISpS rather than ISS.

6. CPF: Delayed information. In this section we study the problem of CPF in the presence of communication delays. We consider the case where all communication channels have the same delay, $\tau > 0$. We further assume that the PF closed-loop subsystems are of Type II, that is, $\eta = \mathbf{0}$.

Motivated by (5.1), we assume that the control law for the reference speed v_{r_i} of each vehicle is given by

$$v_{r_i} = v_L - k_i \sum_{j \in N_{i,p(t)}} (\gamma_i(t) - \gamma_j(t - \tau)). \tag{6.1}$$

Using (2.8) and (6.1), the closed-loop coordination subsystem can be written as

$$\dot{\gamma}(t) = \mathbf{v}_L(\gamma, t) - KD_{p(t)}\gamma(t) + KA_{p(t)}\gamma(t - \tau), \tag{6.2}$$

where D_p and A_p are the degree matrix and the adjacency matrix of the communication graph, respectively. We now determine conditions under which coordination is achieved, that is, under which there exists a signal $\gamma_0(t)$ such that $\gamma = \gamma_0(t)\mathbf{1}$ is a solution of (6.2). Should such a solution exist, then substituting it in (6.2) and using the fact that $A_p = D_p - L_p$ yields

$$\dot{\gamma}_0\mathbf{1} = v_L(\gamma_0, t)\mathbf{1} - KD_p\gamma_0(t)\mathbf{1} + K(D_p - L_p)\gamma_0(t - \tau)\mathbf{1},$$

which simplifies to

$$\dot{\gamma}_0\mathbf{1} - v_L\mathbf{1} = -(\gamma_0(t) - \gamma_0(t - \tau))KD_p\mathbf{1}. \tag{6.3}$$

The equality above is verified iff all elements of the right-hand side vector are equal. For this to be true, one of the following two conditions must apply:

[C1] $\gamma_0(t)$ is either a constant or a periodic signal with period τ. In this case $\gamma_0(t) - \gamma_0(t-\tau) = 0 \; \forall t$ and (6.3) holds with $\dot{\gamma}_0 = v_L$. This condition is not relevant from a practical standpoint.

[C2] $\forall t$, $KD_{p(t)} = kI$ for some $k > 0$. This requires that the degrees of the nodes of the switching communication graph $\mathcal{G}_p$ never vanish, that is, $|N_{i,p}| \neq 0 \; \forall t$, so that the degree matrix D_p is always nonsingular and we can set the control gains to $K = kD_p^{-1}$. In this case, the control gains become piecewise constant functions of p.

In view of the above discussion we consider only condition C2. To lift the constraint $|N_{i,p}| \neq 0$ and have the CPF algorithm be applicable to more general types of switching topologies, we will later modify the control law (6.1). In what follows, we assume v_L is constant. We start by studying the convergence properties of only the coordination dynamics in Lemmas 6.1 and 6.2 below. This is followed by the analysis of the combined PF and coordination systems in Theorem 6.3.

LEMMA 6.1. *Consider the coordination system dynamics* (2.8) *with the control law* (6.1). *Assume that $|N_{i,p(t)}| \neq 0 \; \forall \; t$, and let the control gains be $k_i(t) = k/|N_{i,p(t)}|$. Then, the states γ_i uniformly exponentially agree if the underlying communication graph $\mathcal{G}_p$ is UCM. In this situation, $|\gamma_i - \gamma_j| \to 0$ and $\dot{\gamma}_i \to \dot{\gamma}_0$ as $t \to \infty$, where γ_0 is a solution of the delay differential equation*

$$\dot{\gamma}_0 = -k(\gamma_0(t) - \gamma_0(t-\tau)) + v_L. \tag{6.4}$$

Proof. As explained before, with the control law (6.1) the coordination system takes the form (6.2). Let

$$\tilde{\gamma}(t) = \gamma(t) - \gamma_0(t)\mathbf{1} \tag{6.5}$$

and substitute γ from (6.5) in (6.2) to obtain

$$\begin{aligned} \dot{\gamma}_0(t)\mathbf{1} + \dot{\tilde{\gamma}} = & -K(t)D_{p(t)}\tilde{\gamma}(t) + K(t)A_{p(t)}\tilde{\gamma}(t-\tau) + \\ & -K(t)D_{p(t)}\gamma_0(t)\mathbf{1} + K(t)A_{p(t)}\gamma_0(t-\tau)\mathbf{1} + v_L\mathbf{1}, \end{aligned} \tag{6.6}$$

which simplifies to

$$\dot{\tilde{\gamma}} = -k\tilde{\gamma}(t) + kD_p^{-1}A_p\tilde{\gamma}(t-\tau) \tag{6.7}$$

if $\gamma_0(t)$ is the solution of (6.4) and $K(t) = kD_p^{-1}$. From Theorem 3.5, states $\tilde{\gamma}_i$ in (6.7) agree uniformly exponentially. In particular, $\tilde{\gamma} \to 0$ as $t \to \infty$. Thus, from (6.5) $\gamma \to \gamma_0\mathbf{1}$, and the results follow. □

In general, if v_L is not constant the delayed differential equation (6.4) has no closed form solution. However, for the particular case of v_L constant, one solution is $\gamma_0(t) = v_L^* t$, where $v_L^* = \frac{v_L}{1+k\tau}$. Notice that due to the transmission delay τ there is a finite error in the speed tracking; that is, $\dot{\gamma}_i$ converges to v_L^* and not to v_L.

Consider now the case where there are instants of t time at which $|N_{i,p(t)}| = 0$ for some $i \in \mathcal{N}$. Notice that with the setup adopted in this paper, this condition will necessarily hold over a countable number of disjoint intervals of time, where the length of each interval is bounded above and below by T_0 and τ_D, respectively.

In this case, (6.2) can be rewritten in terms of $\tilde{\gamma}$ defined in (6.5) as

$$\dot{\tilde{\gamma}} = -K(t)D_{p(t)}\tilde{\gamma}(t) + K(t)A_{p(t)}\tilde{\gamma}(t-\tau) + v_L^*\tau(kI - K(t)D_p)\mathbf{1}. \tag{6.8}$$

Clearly, when $\tau = 0$ agreement is achieved for any choice of positive definite K due to Theorem 3.5. However, this is not necessarily the case when $\tau \neq 0$. To see this, assume, for example, that the agreement dynamics (6.8) are at rest, that is, $\dot{\tilde{\gamma}}_i = 0$ $\forall i \in \mathcal{N}$. Then, if $|N_{i,p(t)}| = 0$ for some i and t in a given interval of time, the dynamics of the ith row of (6.8) become $\dot{\tilde{\gamma}}_i = v_L - v_L^*$. This problem can be resolved by applying different desired speeds when vehicle i has no neighbors. The solution is stated next.

Lemma 6.2. *Consider the coordination system dynamics with control law*

$$v_{r_i} = \begin{cases} v_L + \frac{k}{|N_{i,p}|} \sum_{j \in N_{i,p}} \gamma_i(t) - \gamma_j(t-\tau), & N_{i,p} \neq \emptyset, \\ v_L^*, & N_{i,p} = \emptyset, \end{cases} \tag{6.9}$$

where $k > 0$. Then, the states γ_i uniformly exponentially agree if the underlying communication graph $\mathcal{G}_p$ is UCM. In this case, $|\gamma_i - \gamma_j| \to 0$ and $\dot{\gamma}_i \to v_L^*$ as $t \to \infty$.

Proof. The closed-loop coordination dynamics can be expressed in vector form as

$$\dot{\gamma} = -KD_p\gamma(t) + KA_p\gamma(t-\tau) + \frac{v_L - v_L^*}{k} KD_p\mathbf{1} + v_L^*\mathbf{1}.$$

Letting $\gamma(t) = v_L^* t\mathbf{1} + \tilde{\gamma}(t)$ simplifies the closed-loop dynamics to

$$\dot{\tilde{\gamma}} = -KD_p\tilde{\gamma}(t) + KA_p\tilde{\gamma}(t-\tau).$$

Theorem 3.5 implies that $\tilde{\gamma}$ and $\dot{\tilde{\gamma}}$ will converge to the span$\{\mathbf{1}\}$ and to $\mathbf{0}$, respectively, as $t \to \infty$. This concludes the proof. □

Notice that in order to implement the control law (6.9) the vehicles need to know the delay τ in order to compute v_L^*. This raises the practical issue of how to estimate τ. This issue is not addressed in this paper. The following theorem concludes this section.

Theorem 6.3 (CPF with delay). *Consider system Σ that is obtained by putting together the n PF subsystems satisfying Assumption 2.2 and the coordination subsystems studied in Lemma 6.1 or 6.2. Then, the complete system Σ is ISS with input u_1. In particular, PF errors $\|e_i\|$ tend to some ball around zero, and the coordination errors $|\gamma_i - \gamma_j|$ and the speed tracking errors $|\gamma_i - v_L^*|$ converge to zero exponentially.*

Proof. Using Lemma 6.1 or 6.2, we conclude that $\tilde{v}_{r_i} = v_{r_i} - v_L = \dot{\gamma}_i - v_L$ converges to $v_L - v_L^*$ exponentially. Close examination of (2.7) shows that the PF and coordination control subsystems form an interconnected cascade system where $\tilde{v}_{r_i}$ is the output of the coordination control (CC) subsystem and the input to the PF subsystems. Since that latter is ISS from input $\tilde{v}_{r_i}$, the results follow. □

The exposition in this section was strongly motivated by previous work on agreement problems for systems with delays. Especially relevant are the results available in [40] and [8], [10] for continuous time and discrete time, respectively. In particular, the results in [40] address the unforced version of (6.2), that is, with $\mathbf{v}_L(\gamma, t) = \mathbf{0}$. The results in this section reformulate those in [40] to the case where the agreement dynamics are forced by $\mathbf{v}_L(\gamma, t)$.

7. Illustrative example.

This section presents an example that illustrates the application of the CPF techniques developed for the control of three AUVs.

7.1. CPF of three underactuated AUVs.

Consider the problem of CPF control of three underactuated AUVs. Vehicle 2 is allowed to communicate with vehicles 1 and 3, but the latter two do not directly communicate between themselves. To simulate losses in the communications, we considered the situation where both links fail 75% of the time, with the failures occurring periodically with a period of 10[sec]. Moreover, the information transmission delay is 5[sec]. Notice that during failures all the links become deactivated. Since in this scenario the valencies of the nodes vanish periodically, we apply the results of Lemma 6.2. In the simulations, we used the control law (6.9) with $k = 0.1[\mathrm{sec}^{-1}]$.

7.1.1. AUV model. Consider an underactuated vehicle modeled as a rigid body subject to external forces and torques. (See [19] for details on vehicle modeling.) Let {I} be an inertial coordinate frame and {B} a body-fixed coordinate frame whose origin is located at the center of mass of the vehicle. The configuration $(R, \mathbf{p})$ of the vehicle is an element of the special Euclidean group $SE(3) := SO(3) \times \mathbb{R}^3$, where $R \in SO(3) := \{R \in \mathbb{R}^{3\times 3} : RR^T = I_3, \det(R) = +1\}$ is a rotation matrix that describes the orientation of the vehicle and maps body coordinates into inertial coordinates, and $\mathbf{p} \in \mathbb{R}^3$ is the position of the origin of {B} in {I}. Denote by $\nu \in \mathbb{R}^3$ and $\omega \in \mathbb{R}^3$ the linear and angular velocities of the vehicle relative to {I} expressed in {B}, respectively. The following kinematic relations apply:

$$\dot{\mathbf{p}} = R\nu, \tag{7.1a}$$

$$\dot{R} = RS(\omega), \tag{7.1b}$$

where

$$S(x) := \begin{bmatrix} 0 & -x_3 & x_2 \\ x_3 & 0 & -x_1 \\ -x_2 & x_1 & 0 \end{bmatrix} \quad \forall x := (x_1, x_2, x_3)^T \in \mathbb{R}^3.$$

We consider underactuated vehicles with dynamic equations of motion of the form

$$\mathbf{M}\dot{\nu} = -S(\omega)\mathbf{M}\nu + f_\nu(\nu, \mathbf{p}, R) + B_1 u_1, \tag{7.2a}$$

$$\mathbf{J}\dot{\omega} = f_\omega(\nu, \omega, \mathbf{p}, R) + B_2 u_2, \tag{7.2b}$$

where $\mathbf{M} \in \mathbb{R}^{3\times 3}$ and $\mathbf{J} \in \mathbb{R}^{3\times 3}$ denote constant symmetric positive definite mass and inertia matrices, respectively. $u_1 \in \mathbb{R}$ and $u_2 \in \mathbb{R}^3$ denote the control inputs, which act upon the system through a constant nonzero vector $B_1 \in \mathbb{R}^3$ and a constant nonsingular matrix $B_2 \in \mathbb{R}^{3\times 3}$, respectively; and $f_\nu(\cdot)$, $f_\omega(\cdot)$ represent all the remaining forces and torques acting on the body. For the special case of an underwater vehicle, $\mathbf{M}$ and $\mathbf{J}$ also include the so-called hydrodynamic added-mass M_A and added-inertia J_A matrices, respectively, i.e., $\mathbf{M} = M_{RB} + M_A$, $\mathbf{J} = J_{RB} + J_A$, where M_{RB} and J_{RB} are the rigid-body mass and inertia matrices, respectively.

A solution to the PF problem (defined in section 2) of an AUV was given in [1], [2], where the control laws require that $\dot{\gamma}_i$ and $\ddot{\gamma}_i$ be known. Recall that we decomposed the desired speed profile into two parts as $v_{r_i} = v_L + \tilde{v}_{r_i}$ in which only the derivatives of v_L can be computed accurately. However, it can be shown that in the control laws of [1], [2], if the terms $\dot{\gamma}_i$ and $\ddot{\gamma}_i$ are replaced with v_L and $\dot{v}_L$, respectively, the resulting PF closed-loop system becomes ISpS from $\tilde{v}_{r_i}$ as an input. This leads to the following result.

THEOREM 7.1 (PF-AUV). *Consider an underactuated AUV with the equations of motion given by* (7.1) *and* (7.2) *and a desired path* $\mathbf{p}_d(\gamma)$ *in* 3*D-space to be followed. There is a control law for* u_1 *and* u_2 *as functions of the local states* $\mathbf{p}_d$ *and* v_L *that makes the closed-loop system satisfy Assumption* 2.2.

Proof. See the appendix. □

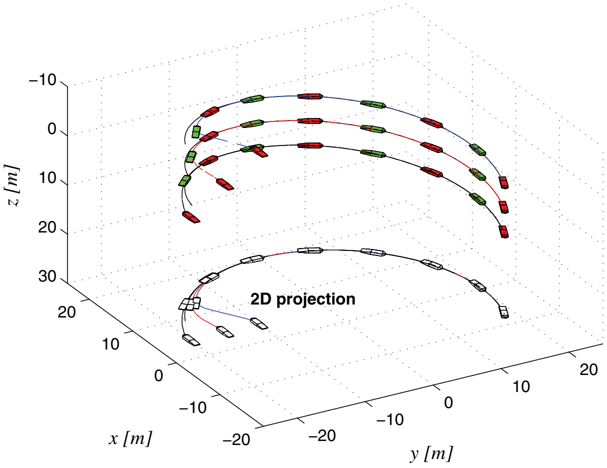


FIG. 4. *CPF of three AUVs, with communication losses and delay.*

Remark 7.2. It is important to notice that the particular PF algorithm that we derive yields input-to-state practical stability, not input-to-state stability. However, the key results obtained in the paper hold true, for input-to-state practical stability can always be viewed as an input-to-state stability condition with an extra constant input.

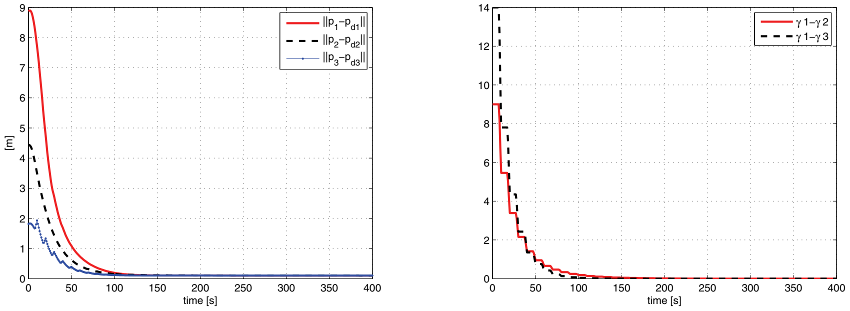
7.1.2. Simulations. In the simulations, the AUVs are required to follow three similar spatial paths shifted along the depth coordinate; that is, the paths are of the form

$$\mathbf{p}_{d_i}(\gamma_i) = \left[c_1 \cos\left(\frac{2\pi}{T}\gamma_i + \phi_d\right), c_1 \sin\left(\frac{2\pi}{T}\gamma_i + \phi_d\right), c_2\gamma_i + z_{0_i}\right],$$

with $c_1 = 20\,\text{m}$, $c_2 = 0.05\,\text{m}$, $T = 400$, $\phi_d = -\frac{3\pi}{4}$ and $z_{0_1} = -10\,\text{m}$, $z_{0_2} = -5\,\text{m}$, $z_{0_3} = 0\,\text{m}$. The initial conditions are $\mathbf{p}_1 = (5\,\text{m}, -10\,\text{m}, -5\,\text{m})$, $\mathbf{p}_2 = (5\,\text{m}, -15\,\text{m}, 0\,\text{m})$, $\mathbf{p}_3 = (5\,\text{m}, -20\,\text{m}, 5\,\text{m})$, $R_1 = R_2 = R_3 = I$, and $v_1 = v_2 = v_3 = \omega_1 = \omega_2 = \omega_3 = \mathbf{0}$. The reference speed v_L was set to $v_L = 0.5[\text{sec}^{-1}]$.

The vehicles are also required to keep a formation pattern that consists of having them aligned along a common vertical line. Figure 4 shows the trajectories of the AUVs. Figure 5 illustrates the evolution of the coordination and PF errors when the communication links fail periodically. Clearly, the vehicles adjust their speeds to meet the formation requirements, and the coordination errors $\gamma_{12} := \gamma_1 - \gamma_2$ and $\gamma_{13} := \gamma_1 - \gamma_3$ converge to zero.

8. Conclusions. This paper addressed the problem of steering a group of vehicles along given paths while holding a desired intervehicle formation pattern (coordinated path-following), all in the presence of *communication losses and time delays*. The solution proposed builds on Lyapunov-based techniques and addresses explicitly the constraints imposed by the topology of the intervehicle communications network. The problem of temporary communication failures was addressed under two scenarios: "brief connectivity losses" and "connected in mean" communication graphs. With the framework adopted, path-following and coordinated control system design become



(a) Path-following errors

(b) Vehicle coordination errors

FIG. 5. 75% *of temporal communication losses; time delay* 5[*sec*].

partially decoupled. As a consequence, the dynamics of each autonomous underwater vehicle can be dealt with by each vehicle controller locally, at the path-following control level. Coordination can then be achieved by resorting to a decentralized control law whereby the exchange of data among the vehicles is kept at a minimum. The system obtained by putting together the path-following and the vehicle coordination strategies proposed was shown to be either a feedback interconnection or a cascade of two input-to-state stable systems. Stability and convergence properties of the resulting interconnected system were studied formally by introducing a new small-gain theorem for systems with brief instabilities. Simulations illustrated the efficacy of the solution proposed.

Further work is required to extend the methodology proposed to tackle more complex coordination control problems, namely, coordinated control in the presence of stringent communication constraints that arise in the underwater world such as nonhomogeneous time variable delays, tight energy budgets, and reduced channel capacity. In particular, the study of coordinated path-following control systems yielding quantifiable measures of performance in the case of unidirectional, event driven communications, is warranted.

Appendix.

Proof of Lemma 3.2.

1. Since $\mathrm{Rank}(I - \mathcal{L}_\beta) = 1$, $\mathcal{L}_\beta$ has $n-1$ eigenvalues at 1. Using the definition of $\mathcal{L}_\beta$, it can be easily verified that $\mathcal{L}_\beta \mathbf{1} = \mathbf{0}$ and $\beta^T \mathcal{L}_\beta = \mathbf{0}^T$, that is, zero is an eigenvalue. Therefore, we can conclude that zero is a single eigenvalue.
2. $\mathcal{L}_\beta K L_p = (K - \frac{1}{\beta^T \mathbf{1}} \mathbf{1}\mathbf{1}^T) L_p = K L_p$, since $\mathbf{1}^T L_p = \mathbf{0}^T$.
3. Straightforward computations show that $\mathcal{L}_\beta^T K^{-1} \mathcal{L}_\beta = K^{-1} - \frac{1}{\beta^T \mathbf{1}} \beta \beta^T$. Therefore, $\nu^T \mathcal{L}_\beta^T K^{-1} \mathcal{L}_\beta \nu = \nu^T K^{-1} \nu - \frac{1}{\beta^T \mathbf{1}} \nu^T \beta \beta^T \nu \leq \nu^T K^{-1} \nu$ for any $\nu \in \mathbb{R}^n$ and the equality holds for $\beta^T \nu = 0$, thus proving the result.
4. The result follows from the fact that $\tilde{\gamma} = \mathcal{L}_\beta \gamma$, $\mathcal{L}_\beta \mathbf{1} = \mathbf{0}$, and $\mathrm{Rank}\, \mathcal{L}_\beta = n-1$.
5. This follows from the definition of $\tilde{\gamma}$ in (3.4).
6. This follows from the definition of $\tilde{\gamma}$ and the fact that $L_p \mathbf{1} = \mathbf{0}$.
7. From

$$|\tilde{\gamma}_i - \tilde{\gamma}_j|^2 = \tilde{\gamma}_i^2 + \tilde{\gamma}_j^2 - 2\tilde{\gamma}_i \tilde{\gamma}_j \leq 2(\tilde{\gamma}_i^2 + \tilde{\gamma}_j^2) \leq 2\|\tilde{\gamma}\|^2 < 2\epsilon^2$$

and $\tilde{\gamma}_i - \tilde{\gamma}_j = \gamma_i - \gamma_j$ it follows that $|\gamma_i - \gamma_j| < \sqrt{2}\epsilon$. Furthermore, from

$KL_p\gamma = KL_p\tilde{\gamma}$ it follows that $\|KL_p\gamma\| \leq \|K\|.\|L_p\|.\|\tilde{\gamma}\| \leq n\epsilon\|K\|$, where we used the fact that $\|L_p\| \leq n$ and equality occurs for a complete graph, that is, for $p = [1, \dots, 1]^T$.

8. Recall the fact that if a graph is connected ($p \in P_c$), then L_p has a single eigenvalue at zero associated to the (right and left) eigenvector $\mathbf{1}$, and the rest of the eigenvalues are positive. Let L be a representative graph Laplacian of L_p for $p \in P_c$. Then, there is a unitary matrix $U = [u_1, \dots, u_n]$ with $u_1 = \frac{1}{\sqrt{n}}\mathbf{1}$ and a diagonal matrix $\Lambda = \operatorname{diag}[\lambda_1, \lambda_2, \dots, \lambda_n]$ with $0 = \lambda_1 < \lambda_2 \leq \cdots \leq \lambda_n$ such that $L = U\Lambda U^T$. For any $\nu \in \mathbb{R}^n$,

$$\begin{aligned}\nu^T L\nu &= \sum_{i=1}^{n} \lambda_i (u_i^T \nu)^2 \\ &= \sum_{i=2}^{n} \lambda_i (u_i^T \nu)^2 \\ &\geq \lambda_2 \sum_{i=2}^{n} (u_i^T \nu)^2 \\ &= \lambda_2 \sum_{i=1}^{n} (u_i^T \nu)^2 - \lambda_2 (u_1^T \nu)^2 \\ &= \lambda_2 \nu^T \nu - \lambda_2 \tfrac{1}{n}(\mathbf{1}^T \nu)^2.\end{aligned}$$

To compute $\lambda_{2,m}$, simply observe that the second term on the right-hand side of the inequality above is zero. Therefore, $\lambda_{2,m}$ is the minimum λ_2 over $p \in P_c$. If $\beta \neq \mathbf{1}$, a standard minimization of the vector function $\nu^T\nu - \frac{1}{n}(\mathbf{1}^T\nu)^2$ with constraints $\beta^T\nu = 0$ and $\nu^T\nu = 1$ yields the results. Similarly, it can be shown that $\bar{\lambda}_m > 0$. Simple numerical computations show that $\lambda_{2,m} = \bar{\lambda}_m$.

9. Recall that the graph Laplacian is $L = D - A$. Using the definitions of degree matrix D and adjacency matrix A, the result follows easily.

10. Because $v_L(\gamma_i, t)$ is bounded and Lipschitz, $|v_L(\gamma_i, t) - v_L(\gamma_j, t)| \leq 2v_M$ and $|v_L(\gamma_i, t) - v_L(\gamma_j, t)| \leq l|\gamma_i - \gamma_j| = l|\tilde{\gamma}_i - \tilde{\gamma}_j| \leq \sqrt{2}l\|\tilde{\gamma}\|$. Then, using

$$\|\mathcal{L}_\beta \mathbf{v}_L(\gamma, t)\|^2 = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} \frac{v_L(\gamma_i, t) - v_L(\gamma_j, t)}{\sigma_j} \right)^2,$$

where $\sigma_j = k_j \sum_i \frac{1}{k_i}$, it is easy to show that $\|\mathcal{L}_\beta \mathbf{v}_L(\gamma, t)\| \leq \sqrt{2n}l\|\tilde{\gamma}\|$ and $\|\mathcal{L}_\beta \mathbf{v}_L(\gamma, t)\| \leq 2\sqrt{n}v_M$, and the result follows.

Proof of Lemma 5.1. First we show that

$$\sum_{i=1}^{n} |\tilde{v}_{r_i}|^2 = \tilde{\gamma}^T L_p K^2 L_p \tilde{\gamma} \leq n^2 k_M^2 \|\tilde{\gamma}\|^2, \tag{A.1}$$

where $\tilde{v}_{r_i}$ and $\tilde{\gamma}$ are as defined in (5.2) and (3.4), respectively. Denote by $l_{i,p}$ the ith column (or row) of L_p. Then $\tilde{v}_{r_i} = k_i l_{i,p}^T \gamma$ and

$$\begin{aligned}\sum_i |\tilde{v}_{r_i}|^2 &= \sum_i k_i^2 \gamma^T l_{i,p} l_{i,p}^T \gamma \\ &= \gamma^T \sum_i k_i^2 l_{i,p} l_{i,p}^T \gamma \\ &= \gamma^T L_p K^2 L_p \gamma \\ &= \tilde{\gamma}^T L_p K^2 L_p \tilde{\gamma}.\end{aligned}$$

Now, consider n PF subsystems, each satisfying Assumption 2.2, and let $\zeta = [\zeta_i]_{n\times 1}$ and $V_1 = \sum_i W_i$. Using (2.6), (2.7), and (A.1) yields

$$\begin{aligned}&\underline{\alpha}_1\|\zeta\|^2 \le V_1 \le \bar{\alpha}_1\|\zeta\|^2,\\ &\dot{V}_1 \le -\lambda_1 V_1 + \rho_1 n^2 k_M^2\|\tilde{\gamma}\|^2 + u_1^2.\end{aligned}$$

Integrating the above differential inequality, the ISS property (5.6) follows.

Proof of Proposition 4.1 (*system interconnection*). Define $V = V_1 + aV_2$ for some $a > 0$ to be chosen later. Clearly, V satisfies the first condition of (4.6) for some $\underline{\alpha} > 0$, $\bar{\alpha} > 0$. Next, we will show that the second condition is also satisfied. Taking the derivative of V yields

$$\dot{V} \le -\left(\lambda_1 - \frac{a\rho_2}{\underline{\alpha}_1}\right)V_1 - a\left(\lambda_2 - \frac{\rho_1}{a\underline{\alpha}_2}\right)V_2 + g\|d\|^2,$$

where $g = \max(1, a)$. At this stage assume ρ_1 and ρ_2 are nonzero, and let

$$\text{(A.2)} \qquad \lambda_0 = \lambda_1 - \frac{a\rho_2}{\underline{\alpha}_1} = \lambda_2 - \frac{\rho_1}{a\underline{\alpha}_2}.$$

Consider the case where $\lambda_2(t) = \lambda_2 > 0$ is constant. If $\rho_1\rho_2 < \underline{\alpha}_1\underline{\alpha}_2\lambda_1\lambda_2$, there exist positive numbers λ_0 and a satisfying (A.2). As a consequence, $\dot{V} \le -\lambda_0 V + g\|d\|^2$, the interconnected system is ISS with input d, and the convergence rate is $\lambda = \lambda_0$.

Consider now the case where $\lambda_2(t)$ is not constant and system 2 has brief instabilities characterized by $\chi(p)$ and $\lambda_2(t)$ as in (4.3). Using the same Lyapunov function $V = V_1 + aV_2$ and λ_0 as in (A.2), compute the derivative of V to obtain

$$\dot{V} \le -\lambda_0 V + a(\lambda_2 - \lambda_2(t))V_2 + g\|d\|^2$$

that yields

$$\dot{V} \le \begin{cases} -\lambda_0 V + g\,\|d\|^2, & \chi(p) = 0,\\ (\lambda_3 - \lambda_0)V + g\,\|d\|^2, & \chi(p) = 1,\end{cases}$$

where $\lambda_3 := \lambda_2 + \tilde{\lambda}_2$. Again, λ_0 exists if $\rho_1\rho_2 < \underline{\alpha}_1\underline{\alpha}_2\lambda_1\lambda_2$. Rewrite

$$\dot{V} \le -\lambda_0 V + a(\lambda_2 - \lambda_2(t))V_2 + g\|d\|$$

and use $aV_2 = V - V_1$ to derive

$$\dot{V} \le -(\lambda_0 - \lambda_3\chi(p))V + g\|d\|^2,$$

where $\lambda_3 := \lambda_2 + \tilde{\lambda}_2$. Integrating the above differential inequalities, it is easy to show that

$$V(t) \le V(t_0)e^{-\lambda_0(t-t_0)+\lambda_3 T_p} + g\sup_{[t_0,t]}\|d\|^2\int_{t_0}^{t} e^{-\lambda_0(t-\tau)+\lambda_3 T_p}d\tau.$$

This yields

$$V(t) \le V(t_0)e^{-(\lambda_0-\alpha\lambda_3)(t-t_0)+\lambda_3 T_\alpha} + \frac{e^{\lambda_3 T_\alpha}}{\lambda_0 - \alpha\lambda_3}g\sup_{[t_0,t]} d^2,$$

where $T_\alpha = (1-\alpha)T_0$ if the system has brief instabilities as defined in (3.3). Therefore, the interconnected system is ISS with d as input if $\alpha < \lambda_0/\lambda_3$.

Suppose now that $\rho_2 = 0$ and $\rho_1 > 0$. In this case, the interconnected system takes a cascade configuration and the dynamics of system 2 are reduced to

$$\dot{V}_2 \leq \begin{cases} -\lambda_2 V_2 + d_2^2, & \chi(p) = 0, \\ \tilde{\lambda}_2 V_2 + d_2^2, & \chi(p) = 1, \end{cases}$$

whose solution takes the form

$$V_2(t) \leq V_2(t_0) e^{-(\lambda_2 - \alpha\lambda_3)(t-t_0)+\lambda_3 T_\alpha} + \frac{e^{\lambda_3 T_\alpha}}{\lambda_2 - \alpha\lambda_3} \sup\nolimits_{[t_0,t]} d_2^2.$$

Using the above inequality together with (4.1) and (4.2) it is easy to obtain

$$V_1(t) \leq a_1 e^{-\lambda_1 t} + a_2 e^{-(\lambda_2-\alpha\lambda_3)t} + a_3 \sup\nolimits_{[t_0,t]} \|d\|^2$$

for some $a_i \geq 0$, $i = 1,2,3$. Therefore, the cascade system is ISS with d as input if $\alpha < \lambda_2/\lambda_3$ and the convergence rate will be $\min(\lambda_1, \lambda_2 - \alpha\lambda_3)$.

Proof of (5.15). The objective is to make $\bar{\lambda}_0 > 0$, that is, $\lambda_\alpha - c_1 l\sqrt{2n} > 0$. Replacing $c_1 = \sqrt{ck_M/k_m}$ in the above inequality yields

$$\lambda_\alpha - l\sqrt{2n} e^{\lambda_\alpha \bar{T}_0} \sqrt{\frac{k_M}{k_m}} > 0.$$

The left-hand side of the inequality takes its maximum at

$$\lambda_\alpha = \frac{1}{\bar{T}_0} \ln\left(\frac{1}{l\sqrt{2n}\bar{T}_0}\sqrt{\frac{k_m}{k_M}}\right),$$

from which it follows that

$$\max \bar{\lambda}_0 = \frac{1}{\bar{T}_0} \ln\left(\frac{1}{el\sqrt{2n}\bar{T}_0}\sqrt{\frac{k_m}{k_M}}\right).$$

To make $\bar{\lambda}_0$ positive it is required that

$$\frac{1}{el\bar{T}_0}\sqrt{\frac{k_m}{2nk_M}} > 1.$$

Using $\bar{T}_0 \leq T + \tau_D$ gives

$$l\sqrt{\frac{k_M}{k_m}} < \frac{1}{e(T+\tau_D)\sqrt{2n}},$$

from which the result follows.

Proof of Theorem 7.1. *PF of an underactuated AUV.* The methodology adopted for PF control system design is rooted in Lyapunov-based and backstepping techniques. The exposition that follows is based on the work in [2].

Step 1. Define the global diffeomorphic coordinate transformation

$$\mathbf{e} := R^T[\mathbf{p} - \mathbf{p}_d(\gamma_i)],$$

which expresses the path tracking error $\mathbf{p} - \mathbf{p}_d$ in a body-fixed frame. For simplicity of presentation, we will for the most part drop the index i in this section. Recall the definition of speed tracking error $\eta = \dot{\gamma}_i - v_r$, where v_r is a reference speed profile. Recall also how the reference speed v_r is decomposed as $v_r = v_L + \tilde{v}_r$, where the derivatives of v_L are known but those of $\tilde{v}_r$ are not. The derivative of $\mathbf{e}$ yields

$$\dot{\mathbf{e}} = -S(\omega)\mathbf{e} + \nu - v_L R^T \mathbf{p}_d^\gamma - \tilde{\eta} R^T \mathbf{p}_d^\gamma,$$

where $\tilde{\eta} := \eta + \tilde{v}_r$ (or equivalently $\dot{\gamma}_i = v_L + \tilde{\eta}$) and superscript γ stands for partial derivative with respect to γ. For example, $\mathbf{p}_d^\gamma = \frac{\partial \mathbf{p}_d}{\partial \gamma}$ and $\mathbf{p}_d^{\gamma^2} = \frac{\partial^2 \mathbf{p}_d}{\partial \gamma^2}$.

We define the Lyapunov function $W_1 := \frac{1}{2}\mathbf{e}^T\mathbf{e}$ and compute its time derivative to obtain

$$\dot{W}_1 = \mathbf{e}^T(\nu - v_L R^T \mathbf{p}_d^\gamma) - \tilde{\eta}\mathbf{e}^T R^T \mathbf{p}_d^\gamma,$$

where we used the fact that $\mathbf{e}^T S(\omega)\mathbf{e} = 0 \; \forall \mathbf{e}, \omega \in \mathbb{R}^3$. We regard ν as a virtual control signal and introduce the virtual control tracking error variable

$$z_1 := \nu - v_L R^T \mathbf{p}_d^\gamma + k_e M^{-1}\mathbf{e}.$$

Then, $\dot{W}_1$ can be rewritten as

$$\dot{W}_1 = -k_e \mathbf{e}^T M^{-1}\mathbf{e} + \mathbf{e}^T z_1 + \alpha_1 \tilde{\eta},$$

where $\alpha_1 := -\mathbf{e}^T R^T \mathbf{p}_d^\gamma$. Ideally, in the absence of $\tilde{\eta}$ one would like to drive z_1 to zero so as to render $\dot{W}_1$ negative. This motivates the next step.

Step 2. The time derivative of z_1 yields

$$M\dot{z}_1 = v_L \Gamma \omega + S(Mz_1)\omega + B_1 u_1 + \tilde{\eta} h_1 + h_2,$$

where

$$\begin{aligned}
\Gamma &:= MS(R^T\mathbf{p}_d^\gamma) - S(MR^T\mathbf{p}_d^\gamma),\\
h_1 &:= -v_L M R^T \mathbf{p}_d^{\gamma^2} - v_L^\gamma M R^T \mathbf{p}_d^\gamma - k_e R^T \mathbf{p}_d^\gamma,\\
h_2 &:= f_\nu + k_e \nu + v_L h_1.
\end{aligned}$$

It turns out that due to lack of actuation, it is not always possible to drive z_1 to zero. Instead, we drive z_1 to a constant design vector $\delta \in \mathbb{R}^3$. To this effect, we define a new error vector $\phi := z_1 - \delta$ and the augmented Lyapunov function

$$W_2 := W_1 + \frac{1}{2}\phi^T M^2,$$

whose derivative is

$$\dot{W}_2 = -k_e \mathbf{e}^T M^{-1}\mathbf{e} + \mathbf{e}^T \delta + \phi^T M(B\zeta + M^{-1}\mathbf{e} + h_2) + \alpha_2 \tilde{\eta},$$

with $\alpha_2 := \alpha_1 + \phi^T M h_1$,

$$B := \begin{pmatrix} B_1 & S(M\delta) + v_L\Gamma \end{pmatrix}, \text{ and } \zeta := \begin{pmatrix} u_1 \\ \omega \end{pmatrix},$$

where we used the fact that $\phi^T MS(Mz_1)\omega = \phi^T MS(M\delta)\omega$. Matrix B can always be made full rank; see [2] for details. Let

$$\beta_1 := B^T(BB^T)^{-1}(-h_2 - M^{-1}\mathbf{e} - k_\phi \phi).$$

To complete this step, we set u_1 to be the first entry of β_1, that is, $u_1 = \begin{pmatrix} 1 & 0_{1\times 3} \end{pmatrix}\beta_1$, and introduce the error variable

$$z_2 := \omega - \Pi\beta_1, \quad \Pi := \begin{pmatrix} 0_{3\times 1} & I_{3\times 3} \end{pmatrix}$$

that should be driven to zero. It follows that

$$\dot{W}_2 = -k_e \mathbf{e}^T M^{-1}\mathbf{e} + \mathbf{e}^T\delta - k_\phi \phi^T M\phi + \phi^T MB\Pi^T z_2 + \alpha_2\tilde{\eta}.$$

Step 3. Let $\dot{\beta}_1 := h_3 + h_4\tilde{\eta}$, where h_3 collects the terms in $\dot{\beta}_1$ not containing $\tilde{\eta}$. For simplicity we do not expand h_3 and h_4. Define

$$W_3 := W_2 + \frac{1}{2}z_2^T J z_2,$$

whose time derivative, after applying the control law

$$u_2 = B_2^{-1}(-f_\omega + J\Pi h_3 - \Pi B^T M\phi - k_z z_2),$$

yields

$$\dot{W}_3 = -k_e \mathbf{e}^T M^{-1}\mathbf{e} + \mathbf{e}^T\delta - k_\phi \phi^T M\phi - k_z z_2^T z_2 + \alpha_3\tilde{\eta}, \tag{A.3}$$

where $\alpha_3 = \alpha_2 - z_2^T J\Pi h_4$. At this point it is important to notice that

$$|\alpha_3| \le k_1\|\mathbf{e}\| + k_2\|\phi\| + k_3\|z_2\| \tag{A.4}$$

for some $k_i > 0$, $i = 1, 2, 3$, that are functions of v_L, v_L^γ, M, $\mathbf{p}_d^\gamma$, and $\mathbf{p}_d^{\gamma^2}$. The design phase is concluded at this step for the case where $\eta = 0$ simply by making $\dot{\gamma}_i = v_r$. In this case, $\tilde{\eta} = \tilde{v}_r$ and

$$\dot{W}_3 \le -\lambda W_3 + \rho_1\|\delta\|^2 + \rho_2|\tilde{v}_r|$$

for some $\lambda > 0$, $\rho_1 > 0$, and $\rho_2 > 0$. That is, the PF closed-loop system is ISpS with input $\tilde{v}_r$, state $x_1 = (\mathbf{e}, \phi, z_2)^T$, and constant $\rho_1\|\delta\|^2$.

Step 4. This extra step contemplates the situation where $\eta \neq 0$. To this effect, augment the Lyapunov function W_3 to obtain

$$W_4 := W_3 + \frac{1}{2}\eta^2 = \frac{1}{2}\mathbf{e}^T\mathbf{e} + \frac{1}{2}\phi^T M^2\phi + \frac{1}{2}z_2^T J z_2 + \frac{1}{2}\eta^2.$$

Set the feedback law

$$\dot{\eta} = -\alpha_3 - k_\eta \eta$$

to make

$$\dot{W}_4 = -k_e \mathbf{e}^T M^{-1}\mathbf{e} - k_\phi \phi^T M\phi - k_z z_2^T z_2 - k_\eta \eta^2 + \mathbf{e}^T\delta + \alpha_3\tilde{v}_r,$$

which can be rewritten as

$$\dot{W}_4 \le -\lambda W_4 + \rho_1\|\delta\|^2 + \rho_2|\tilde{v}_r| \tag{A.5}$$

for some $\lambda > 0$, $\rho_1 > 0$, and $\rho_2 > 0$. Again, this makes the closed-loop system ISpS with input $\tilde{v}_r$, state $x_1 = (\mathbf{e}, \phi, z_2, \eta)^T$, and constant $\rho_1\|\delta\|^2$.

REFERENCES

[1] A. P. Aguiar and J. P. Hespanha, *Logic-based switching control for trajectory-tracking and path-following of underactuated autonomous vehicles with parametric modeling uncertainty*, in Proceedings of the American Control Conference, Boston, MA, 2004, pp. 3004–3010.

[2] A. P. Aguiar and J. P. Hespanha, *Trajectory-tracking and path-following of underactuated autonomous vehicles with parametric modeling uncertainty*, IEEE Trans. Automat. Control, 52 (2007), pp. 1362–1379.

[3] A. P. Aguiar, J. P. Hespanha, and P. Kokotovic, *Path-following for non-minimum phase systems removes performance limitations*, IEEE Trans. Automat. Control, 50 (2005), pp. 234–239.

[4] A. P. Aguiar, I. Kaminer, R. Ghabcheloo, A. M. Pascoal, N. Hovakimyan, C. Cao, and V. Dobrokhodov, *Coordinated path following of multiple UAVs for time-critical missions in the presence of time-varying communication topologies*, in Proceedings of the 17th IFAC World Congress, Korea, 2008, pp. 16015–16020.

[5] M. Arkak, *Passivity as a design tool for group coordination*, IEEE Trans. Automat. Control, 52 (2007), pp. 1380–1390.

[6] R. Beard, J. Lawton, and F. Hadaegh, *A coordination architecture for spacecraft formation control*, IEEE Trans. Control Systems Tech., 9 (2001), pp. 777–790.

[7] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Prentice–Hall, Englewood Cliffs, NJ, 1989.

[8] V. D. Blondel, J. M. Hendrickx, A. Olshevsky, and J. N. Tsitsiklis, *Convergence in multiagent coordination, consensus, and flocking*, in Proceedings of the 44th IEEE CDC-ECC, Seville, Spain, 2005, pp. 2996–3000.

[9] M. Cao, A. S. Morse, and B. D. O. Anderson, *Agreeing asynchronously: Announcement of results*, in Proceedings of the 46th IEEE Conference on Decision and Control, San Diego, CA, 2006, pp. 4301–4306.

[10] M. Cao, A. S. Morse, and B. D. O. Anderson, *Reaching an agreement using delayed information*, in Proceedings of the 46th IEEE Conference on Decision and Control, San Diego, CA, 2006, pp. 3375–3380.

[11] M. Cao, A. S. Morse, and B. D. O. Anderson, *Reaching a consensus in a dynamically changing environment: Convergence rates, measurement delays, and asynchronous events*, SIAM J. Control Optim., 47 (2008), pp. 601–623.

[12] J. Cortes and F. Bullo, *Coordination and geometric optimization via distributed dynamical systems*, SIAM J. Control Optim., 44 (2005), pp. 1543–1574.

[13] T. B. Curtin, J. G. Bellingham, J. Catipovic, and D. Webb, *Autonomous oceanographic sampling network*, Oceanography, 6 (1993), pp. 86–95.

[14] D. V. Dimarogonas, S. G. Loizou, K. J. Kyriakopoulos, and M. M. Zavlanos, *A feedback stabilization and collision avoidance scheme for multiple independent non-point agents*, Automatica, 42 (2006), pp. 229–243.

[15] W. B. Dunbar and R. M. Murray, *Distributed receding horizon control for multi-vehicle formation stabilization*, Automatica, 42 (2006), pp. 549–558.

[16] M. Egerstedt and X. Hu, *Formation constrained multi-agent control*, IEEE Trans. Robotics and Automation, 17 (2001), pp. 947–951.

[17] P. Encarnação and A. Pascoal, *Combined trajectory tracking and path following: An application to the coordinated control of marine craft*, in Proceedings of the 40th IEEE Conference on Decision and Control (CDC), Orlando, FL, 2001, pp. 964–969.

[18] A. Fax and R. Murray, *Information flow and cooperative control of vehicle formations*, IEEE Trans. Automat. Control, 49 (2004), pp. 1465–1476.

[19] T. Fossen, *Guidance and Control of Ocean Vehicles*, John Wiley & Sons, New York, 1994.

[20] R. Ghabcheloo, A. Aguiar, A. Pascoal, and C. Silvestre, *Synchronization in multi-agent systems with switching topologies and non-homogeneous communication delays*, in Proceedings of the 46th IEEE Conference on Decision and Control (CDC), New Orleans, LA, 2007, pp. 2327–2332.

[21] R. Ghabcheloo, A. Aguiar, A. Pascoal, C. Silvestre, I. Kaminer, and J. Hespanha, *Coordinated path-following control of multiple underactuated autonomous vehicles in the presence of communication failures*, in Proceedings of the 45th IEEE Conference on Decision and Control (CDC), San Diego, CA, 2006, pp. 4345–4350.

[22] R. Ghabcheloo, A. Pascoal, C. Silvestre, and I. Kaminer, *Nonlinear coordinated path-following control of multiple wheeled robots with bi-directional communication constraints*, Internat. J. Adapt. Control Signal Process., 21 (2007), pp. 133–157.

[23] F. Giuletti, L. Pollini, and M. Innocenti, *Autonomous formation flight*, IEEE Control Systems Mag., 20 (2000), pp. 33–34.

[24] C. Godsil and G. Royle, *Algebraic Graph Theory*, Grad. Texts in Math., Springer-Verlag, New York, 2001.

[25] J. M. Hespanha, O. A. Yakimenko, I. I. Kaminer, and A. M. Pascoal, *Linear parametrically varying systems with brief instabilities: An application to vision/inertial navigation*, IEEE Trans. Aerospace Electron. Systems, 40 (2004), pp. 889–900.

[26] I.-A. F. Ihle, *Coordinated Control of Marine Craft*, Ph.D. thesis, Norwegian University of Science and Technology, Trondheim, Norway, 2006.

[27] H. Ito, *A constructive proof of ISS small-gain theorem using generalized scaling*, in Proceedings of the 41st IEEE Conference on Decision and Control, Las Vegas, NV, 2002, pp. 2286–2291.

[28] A. Jadbabaie, J. Lin, and A. S. Morse, *Coordination of groups of mobile autonomous agents using nearest neighbor rules*, IEEE Trans. Automat. Control, 48 (2003), pp. 988–1001.

[29] I. Kaminer, O. Yakimenko, V. Dohrokhodov, A. Pascoal, N. Hovakimyan, V. Patel, C. Cao, and A. Young, *Coordinated path following for time-critical missions of multiple UAVs via L1 adaptive output feedback controllers*, in Proceedings of the AIAA Guidance, Navigation and Control Conference and Exhibit, Hilton Head, SC, Aug. 20–23, 2007, paper AIAA-2007-6409.

[30] I. Kaminer, O. Yakimenko, A. Pascoal, and R. Ghabcheloo, *Path generation, path following and coordinated control for time-critical missions of multiple UAVs*, in Proceedings of the American Control Conference (ACC), Minneapolis, MN, 2006, pp. 4906–4913.

[31] H. K. Khalil, *Nonlinear Systems*, 3rd ed., Prentice–Hall, Englewood Cliffs, NJ, 2002.

[32] Y. Kim and M. Mesbahi, *On maximizing the second smallest eigenvalue of state-dependent graph Laplacian*, IEEE Trans. Automat. Control, 51 (2006), pp. 116–120.

[33] D. J. Klein, C. Matlack, and K. A. Morgansen, *Cooperative target tracking using oscillator models in three dimensions*, in Proceedings of the American Control Conference, New York, 2007, pp. 2569–2575.

[34] E. Kyrkjebo, *Motion Coordination of Mechanical Systems: Leader-Follower Synchronization of Euler-Lagrange Systems Using Output Feedback Control*, Ph.D. thesis, Norwegian University of Science and Technology, Trondheim, Norway, 2007.

[35] L. Lapierre, D. Soetanto, and A. Pascoal, *Coordinated motion control of marine robots*, in Proceedings of the 6th IFAC Conference on Manoeuvring and Control of Marine Craft (MCMC), Girona, Spain, 2003.

[36] J. R. T. Lawton, R. W. Beard, and B. J. Young, *A decentralized approach to formation maneuvers*, IEEE Trans. Robotics and Automation, 19 (2003), pp. 933–942.

[37] Z. Lin, *Coupled Dynamic Systems: From Structure Towards Stability and Stabilizability*, Ph.D. thesis, University of Toronto, Toronto, ON, Canada, 2006.

[38] Z. Lin, B. Francis, and M. Maggiore, *State agreement for continuous-time coupled nonlinear systems*, SIAM J. Control Optim., 46 (2007), pp. 288–307.

[39] M. Mesbahi and F. Hadaegh, *Formation flying control of multiple spacecraft via graphs, matrix inequalities and switching*, J. Guidance Control Dynam., 24 (2001), pp. 369–377.

[40] L. Moreau, *Stability of continuous-time distributed consensus algorithm*, in Proceedings of the 43rd IEEE Conference on Decision and Control, Atlantis, Paradise Island, Bahamas, 2004, pp. 3998–4003.

[41] N. Moshtagh and A. Jadbabaie, *Distributed geodesic control laws for flocking of nonholonomic agents*, IEEE Trans. Automat. Control, 52 (2007), pp. 681–686.

[42] R. M. Murray, *Recent research in cooperative control of multi-vehicle systems*, ASME J. Dynam. Systems Measurement Control, 129 (2007), pp. 571–583.

[43] S. Nair and N. E. Leonard, *Stable synchronization of mechanical system networks*, SIAM J. Control Optim., 47 (2008), pp. 661–683.

[44] R. Olfati-Saber, J. A. Fax, and R. M. Murray, *Consensus and cooperation in networked multi-agent systems*, Proc. IEEE, 95 (2007), pp. 215–233.

[45] A. Olshevsky and J. N. Tsitsiklis, *On the nonexistence of quadratic Lyapunov functions for consensus algorithms*, IEEE Trans. Automat. Control, 53 (2008), pp. 2642–2645.

[46] A. Papachristodoulou and A. Jadbabaie, *Synchronization in oscillator networks: Switching topologies and non-homogeneous delays*, in Proceedings of the 44th IEEE Conference on Decision and Control (CDC), Seville, Spain, 2005, pp. 5692–5697.

[47] A. Pascoal et al., *Robotic ocean vehicles for marine science applications: The European ASIMOV project*, in Proceedings of the OCEANS MTS/IEEE Conference, Providence, RI, 2000, pp. 409–415.

[48] Q. C. Pham and J. J. E. Slotine, *Stable concurrent synchronization in dynamic system networks*, Neural Networks, 20 (2007), pp. 62–77.

[49] M. Pratcher, J. D'Azzo, and A. Proud, *Tight formation control*, J. Guidance Control Dynam., 24 (2001), pp. 246–254.

[50] A. Rahmani, M. Mesbahi, and F. Y. Hadaegh, *On the optimal balanced-energy formation flying maneuvers*, AIAA J. Guidance Control Dynam., 29 (2006), pp. 1395–1403.

[51] W. Ren and R. W. Beard, *Distributed Consensus in Multi-Vehicle Cooperative Control*, Comm. Control Engrg. Ser., Springer-Verlag, New York, 2007.

[52] W. Ren and N. Sorensen, *Distributed coordination architecture for multi-robot formation control*, J. Robotics and Autonomous Systems, 56 (2008), pp. 324–333.

[53] W. J. Rugh, *Linear System Theory*, Prentice–Hall, Englewood Cliffs, NJ, 1993.

[54] R. Sepulchre, D. A. Paley, and N. E. Leonard, *Stabilization of planar collective motion with all-to-all communication*, IEEE Trans. Automat. Control, 52 (2007), pp. 811–824.

[55] D. D. Šiljak, *Decentralized Control of Complex Systems*, Academic Press, Boston, MA, 1991.

[56] D. D. Šiljak, *Dynamic graphs*, Nonlinear Anal. Hybrid Systems, 2 (2008), pp. 544–567.

[57] R. Skjetne, T. I. Fossen, and P. Kokotovic, *Robust output maneuvering for a class of nonlinear systems*, Automatica, 40 (2004), pp. 373–383.

[58] R. Skjetne, S. Moi, and T. Fossen, *Nonlinear formation control of marine craft*, in Proceedings of the 41st IEEE Conference on Decision and Control, Las Vegas, NV, 2002, pp. 1699–1704.

[59] E. D. Sontag and Y. Wang, *New characterizations of input-to-state stability*, IEEE Trans. Automat. Control, 41 (1996), pp. 1283–1294.

[60] D. Stilwell and B. Bishop, *Platoons of underwater vehicles*, IEEE Control Systems Mag., 20 (2000), pp. 45–52.

[61] J. N. Tsitsiklis and M. Athans, *Convergence and asymptotic agreement in distributed decision problems*, IEEE Trans. Automat. Control, 29 (1984), pp. 42–50.

[62] C. Yu, J. M. Hendrickx, B. Fidan, and B. D. O. Anderson, *Three and higher dimensional autonomous formations: Rigidity, persistence and structural persistence*, Automatica, 43 (2007), pp. 387–402.

[63] F. Zhang, D. M. Fratantoni, D. Paley, J. Lund, and N. E. Leonard, *Control of coordinated patterns for ocean sampling*, Internat. J. Control, 80 (2007), pp. 1186–1199.

[64] F. Zhang and P. S. Krishnaprasad, *Co-ordinated orbit transfer of satellite clusters*, Astrodynamics, Space Missions and Chaos (Annals of the New York Academy of Sciences), 1017 (2004), pp. 112–137.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 1, pp. 266–291

© 2009 Society for Industrial and Applied Mathematics

ESTIMATION OF SPATIALLY DISTRIBUTED PROCESSES USING MOBILE SPATIALLY DISTRIBUTED SENSOR NETWORK*

MICHAEL A. DEMETRIOU† AND ISLAM I. HUSSEIN†

Abstract. The problem of estimating a spatially distributed process described by a partial differential equation (PDE), whose observations are contaminated by a zero mean Gaussian noise, is considered in this work. The basic premise of this work is that a set of mobile sensors achieve better estimation performance than a set of immobile sensors. To enhance the performance of the state estimator, a network of sensors that are capable of moving within the spatial domain is utilized. Specifically, such an estimation process is achieved by using a set of spatially distributed mobile sensors. The objective is to provide mobile sensor control policies that aim to improve the state estimate. The metric for such an estimate improvement is taken to be the expected state estimation error. Using different spatial norms, two guidance policies are proposed. The current approach capitalizes on the efficient filter gain design in order to avoid intense computational requirements resulting from the solution to filter Riccati equations. Simulation studies implementing and comparing the two proposed control policies are provided.

Key words. spatially distributed systems, sensor control, mobile sensor network, process estimation, diffusion equation

AMS subject classifications. 93E10, 93C20, 70B05, 93D05

DOI. 10.1137/060677884

1. Introduction and examples.

Many applications have emerged in recent years that rely on the use of networks of dynamic multiagent, limited-range sensors to collect and process data. Applications include emergency response, aerial mapping, and multiple satellite imaging systems for high-resolution imaging [71]. These and other applications often involve tasks in adversarial, highly dynamic environments that are hazardous to human operators. Hence, there is a pressing need to develop autonomous multiagent sensor network systems that seek to collect and process distributed information under constrained resources.

Sensors of interest, such as infrared and vision-based cameras, and sonar, are used to measure a certain field over some domain $\mathcal{D}$. In many applications this is known as the coverage problem. In these problems, the field to be measured satisfies a partial differential equation (PDE). In addition to the task of collecting field measurements, the sensors may also be asked to relay information to the base station or to process the information for ensuing decision making. Data processing may be carried out either in a centralized or decentralized fashion in order to possibly (i) estimate the state, (ii) identify the process parameters, (iii) detect sources, and, in the event of an actuation capability, (iv) provide action in order to alter the process response. Examples include the estimation of the temperature distribution in a wildfire, where the PDE is given by a reaction-convection-diffusion equation [3, 39, 63]. The present work attempts to address the basic goal of using a network of dynamic, limited-range sensors to improve the estimate of the field of interest.

*Received by the editors December 17, 2006; accepted for publication (in revised form) September 2, 2008; published electronically February 11, 2009.
http://www.siam.org/journals/sicon/48-1/67788.html

†Department of Mechanical Engineering, Worcester Polytechnic Institute, Worcester, MA 01609 (mdemetri@wpi.edu, ihussein@wpi.edu). The first author gratefully acknowledges the support from the National Science Foundation, grant CMS-0408974, while the second author gratefully acknowledges financial support from WPI Faculty Development Grant.

The theory and application of network control have recently received much attention, as is evident from the plethora of published works in the last few years. In [59], the author provides a lucid overview of the theory, operation, and application of wireless sensor networks. We also refer the reader to [5], where the authors provide a thorough account of the state of the art and of the current challenges in networked real-time systems. Within the same special issue of *Proceedings of the IEEE*, several papers address various issues such as the current state of the technology of networked control systems, the foundations of networked real-time systems, and wireless networks.

In general, there are two classes of sensor coverage control problems. The first class involves spatially *fixed* sensors. The goal, which has been extensively studied in the past, is to optimize sensor locations and sensor domains in fixed-sensor networks, and the problems in this class are considered to be in locational optimization [37, 73]. In such problems, the solution is a Voronoi partition [38], where the optimal sensor domain is a Voronoi cell in the partition and the optimal sensor location is a centroid of a Voronoi cell in the partition.

The second class of problems involves a set of *mobile* sensors. In [14], the authors present a survey of recent activities in the control and design of both static and dynamic sensor networks. In their design criteria, they consider issues such as maximum coverage, detection of events, and minimum communication energy expenditure. In the paper [65], the authors propose a formal model for a network of robotic agents and define notions of network, control, and communication law; coordination; and time and communication complexity. In a subsequent publication, the authors provide upper and lower bounds on the time complexity of basic coordination algorithms running on synchronous robotic networks for rendezvous and deployment over a region of interest [66].

The authors in [41] discuss challenges in the modeling of robotic networks, motion coordination algorithms, sensing and estimation tasks, and complexity of distributed algorithms. In [67], the authors present recent theoretical tools for modeling, analysis, and design of coordination algorithms for networks of mobile autonomous agents for problems with distributed information. The authors discuss motion coordination and the motivation that some recent techniques has received from biological systems. These problems include deployment over a given region, pattern formation, rendezvous, or synchronous rigid-body motions.

In [60], the authors consider a probabilistic network model and a density function to represent the frequency of random events taking place over a mission space. The authors develop an optimization problem that aims to maximize coverage using sensors with limited ranges, while minimizing communication cost. Starting with initial sensor positions, the authors develop a gradient algorithm to converge to a (local) solution to the optimization problem. The sequence of sensor distributions along the solution is seen as a discrete time trajectory of the mobile sensor network until it converges to the local minimum. In [19], the authors address the same question, but instead of converging to a local solution of some optimization problem, the trajectory converges to the centroid of a cell in a Voronoi partition of the search domain. The authors propose stable algorithms in both continuous and discrete time. These algorithms are the dynamic version of the Lloyd algorithm [62], which iteratively achieves the optimal configuration. Voronoi-based approaches, however, require exhaustive computational effort to compute the Voronoi cells continuously during a real-time implementation of the controllers.

In [18], the authors present coordination algorithms for groups of mobile agents performing deployment and coverage tasks under the constraint that each mobile agent has a limited sensing or communication radius. Based on the geometry of Voronoi partitions and proximity graphs, they propose coverage algorithms in continuous and discrete time that are convergence-guaranteed and are spatially distributed with respect to appropriate proximity graphs. In [40], the authors propose a nonsmooth gradient algorithm for the problem of maximizing the area of the region visible to an observer in a simple nonconvex polygon.

In [64], the authors study optimal sensor placement and motion coordination strategies for mobile sensor networks for target tracking using range sensors. They propose motion coordination algorithms that achieve optimal deployment. The authors in [84] propose an algorithm for monitoring an environmental boundary with mobile agents that use only locally sensed information. Their objective is to approximate the boundary with a polygon. The algorithm proves to be convergent for static boundaries and is shown to perform well for slowly moving boundaries.

The paper [42] uses a novel discrete-event controller for the coordination of cooperating heterogeneous wireless sensor networks containing both unattended ground sensors and mobile sensor robots. Given an environment perception, the discrete-event controller sequences the most suitable tasks for each agent and assigns sensor resources. The authors introduce several new tools for discrete-event controller design and operation. The resulting controller represents a complete dynamical description of the wireless network system and is experimentally demonstrated on a wireless sensor network prototyping system.

In [44], the authors use a stochastic approach to find the sensor schedule that results in the minimum error covariance of a state to be measured. They develop a stochastic sensor selection strategy that is computationally tractable. Applications of this work include the sensor selection problem, where multiple sensors cannot operate simultaneously, as in single frequency band sonar in which sensor trajectory optimization is needed to optimize their trajectories. The algorithm is applied to these problems and illustrated through simple examples.

In the above works, the authors address the *redeployment* problem to improve network performance. More recent research results, such as [50, 51], consider the following problem. Given a sensor network and a mission domain (the domain to be sampled) $\mathcal{D}$, develop closed-loop control strategies such that each point in $\mathcal{D}$ is sampled by some agents in the network by an amount of effective coverage equal to C^*. In the discrete setting, the goal may be understood as the collection of at least C^* measurements of a physical quantity at each point in $\mathcal{D}$ using a group of limited-range sensors. The goal is to dynamically survey the mission domain while the agents are moving in the mission space. This problem is known in the robotics literature as the coverage path planning problem, where a single limited-range sensor agent needs to visit all points in the environment (see, for example [1, 16], and references therein).

For the coverage path planning problem for networks of multiple sensor-equipped robots, in [50] a deterministic approach is pursued and a convergent cooperative feedback control law is proposed that achieves a satisfactory coverage of $\mathcal{D}$, while avoiding converging to local minima of a defined coverage error. These results were motivated by approaches studied in [15, 47] for (optimal and suboptimal) motion planning of multiple spacecraft interferometric imaging systems (MSIIS). In [51], the authors also guarantee collision avoidance, and in [52] they further modify the control law to guarantee collision-free coverage with a flocking behavior.

Regarding collision avoidance, the authors in [51, 52] mainly rely on the use of barrier-type functions originally developed in [57, 58] for collision avoidance in two-agent systems. Later these results were generalized for multiagent (more than two) systems in noncooperative [83] and cooperative settings [82]. A decentralized scheme for collision avoidance of multiple independent nonpoint agents was developed in [36] using a methodology based on navigation functions. In this paper, we do not include collision avoidance control. However, once cooperative coverage control strategies have been developed (as in this paper), one can easily append collision avoidance control components to the coverage control law to achieve safe coverage of the domain.

In the stochastic setting, the authors in [17] develop a Kalman filter-based algorithm that aims to use a mobile sensor network for estimating the state of a *single* target (that is, this is not a coverage control problem). The author in [48] uses the Kalman filter for estimating a spatially decoupled (i.e., it does not satisfy a PDE) field and uses the prediction step of the filter for guiding the vehicles to move in directions that improve the field estimate. Moreover, the control algorithm is modified to guarantee satisfactory global coverage of the domain. A similar coverage problem formulation is addressed in [43], but from an information theoretic perspective that does not discuss dwelling into local minima of the metric of choice. Motivated by the employment of sensing devices for the estimation of spatially distributed processes (unsteady diffusion-type parabolic PDEs) in [26], the authors in [49] extend these results to the case where the field of interest satisfies a PDE. The present paper is an extension of [49], where we presently formalize the mathematical approach and address some of the mathematical intricacies in the formulation. We also present the estimation problem in a more general abstract formulation that is important to understand. Such an abstract framework is conducive to optimization that emanates from the subsequent inclusion of communication and decision/actuation considerations.

A common aspect of all the results mentioned above is the assumption that the field to be measured is static, especially in the spatial sense. In this paper we consider the case where the field to be measured satisfies some PDE, i.e. *evolves both in time and space*. Applications for both immobile and mobile sensor networks include the following:

- Aerial wildfire control in inaccessible and rugged country, where the temperature distribution, satisfying a PDE [3, 39, 63], has to be estimated to identify critical points that require immediate deposition of fire suppression material.
- Underwater and atmospheric sampling, where the field to be measured (e.g., salinity or temperature) satisfies a particular PDE [9, 54, 78, 79].
- Health monitoring of civil infrastructures such as bridges and "smart" buildings [13], wherein a network of sensors is used to monitor vital structural changes due to wear, fire, oxidation, cyclic loading, and earthquakes.
- Oil spill and ground water contamination, where mobile agents with sensing and possible actuating capabilities are used to contain/encircle moving boundaries of contaminating fluids. For the former, flotillas equipped with computational, sensing, and limited actuation capabilities attempt to encircle and contain moving boundaries caused by contaminating oil spills, relay information to a base station, and possibly take limited action by minimizing the environmental effect of the contaminating substance via an appropriate dispersion of neutralizing agents.
- Other applications, such as MSIIS, surveillance, and aerial mapping, where the PDE does not have a diffusion term (since "information" does not diffuse spatially) and the PDE is elliptic. Basically, these are Poisson-like PDEs which describe the steady-state solutions of unsteady diffusion-advection processes. This is a special subclass of PDEs that also fall under the more general

class of PDEs considered here. When such elliptic PDEs are considered, one simply embeds them in an unsteady or parabolic process and proceeds with the guidance scheme proposed in this paper. The embedding of such elliptic PDEs is similar in spirit to the work in [45] for the adaptive parameter identification in ground water hydrology.

Communications, in particular wireless communication, in networks is a crucial issue in cooperative networks. Issues include lossy communications, fading channels, dynamic communication structures, and communication-induced time delays. While this paper does not address the issue of communications aspects, it deserves much attention and is the subject of current and future research by the authors. However, this paper lays down the abstract mathematical framework in which one can naturally augment communication and actuation aspects via the penalization of an associated performance measure. For more on communication in networked systems, we refer the reader to the review paper [5].

In the context of sensor networks, the papers [69, 70] address distributed sensing and communication under communication constraints and derive tradeoffs between communication and sensing requirements in a decentralized mobile sensor network, respectively. We refer the reader to these papers and references therein for more on constraints imposed by communications (such as channel fading) on sensor network performance.

In a parallel fashion to the above research efforts on sensor and actuator networks, there was a considerable amount of research done by the infinite dimensional systems community, which considered the placement and scheduling of sensing and actuating devices in systems governed by spatially distributed processes; such examples include thermal manufacturing, chemical transport processes, and mechanical structures. The basic idea was to optimally place and schedule such devices within the spatial domain. The PDE interpretation on "move" or "schedule" or "scan" sensors and actuators to improve performance of the filter or the controller and to enhance the identifiability of the parameter estimation scheme translated into studying the well-posedness of an associated evolution equation. More specifically, the placement and scheduling of sensing and actuating devices was equivalent to choosing the output and input operators that were parameterized by the spatial position of these devices. An added dimension to the positioning of sensing and actuating devices in processes described by PDEs was the issue of locations that resulted in partial or complete loss of observability and controllability. For a one-dimensional diffusion equation this amounts to the avoidance of placing pointwise-in-space sensing and actuating devices at the zeros of the associate spatial operator of the system; this then relates to the definition of approximate observability and controllability [21]. Such works started to appear, at least in the open literature, in the late 1970s and throughout the 1980s both in the West and the former Soviet Union. For the former, works in [7, 20, 55, 68, 56] provided the seed for viewing the spatial location of sensing and actuating devices as another level of control and optimization. For the latter, early works by Butkovskiĭ [10, 11, 12] paved the way for the eventual guidance of sensors and actuators in systems governed by PDEs and whose state is a field over a spatial domain.

The process to be estimated is naturally described by a PDE. A system-theoretic approach to studying PDEs has received much attention over the years, addressing various issues such as control and filter design, optimization, and finite dimensional approximation. Such efforts have been reported, for example, in the texts [21, 61, 74]. Related to the work under consideration is the issue of the placement of sensors at fixed positions for improved state and parameter estimation, fault tolerance, and observer-based closed-loop performance [20, 29, 86]. Closer to the proposed issue of mobile sensors is the work by Uciński and coworkers [81, 87, 88] and Nehorai and coworkers [72]. The above works dealt primarily with optimal motion planning of mobile sensors (robots) for parameter estimation. By utilizing an information theoretic approach, via the use of the Fisher information matrix, in the optimization of an objective function along with additional robot motion constraints, an optimal path planning policy was derived. While similar to the previous works, the current work considers the motion planning of mobile sensors in distributed systems for improved state estimation.

The proposed mobile sensor motion planning is based on Lyapunov stability techniques. In addition to the method used to derive the motion planning, the proposed work considers spatially distributed sensors as opposed to point sensors. Related to the above is the work in [26], where a network of fixed-position pointwise sensors was utilized for detection of a moving source (intrusion detection) for a diffusion process (i.e., the field to be measured was the solution to a diffusion-advection PDE) in a two-dimensional spatial domain. A sensor management scheme was proposed in order to minimize power consumption by having a subset of the available sensors in transmit mode and having the remaining sensors in the network in sleep mode. The detection scheme would activate, over the duration of a given time interval, the relevant sensors within a radius to the moving source and deactivate the sensors that were outside a ball surrounding the centroid of the moving source. A state estimator was subsequently incorporated into the moving source detection scheme in [28] for the same diffusion process, and eventually a containment policy utilizing actuating devices collocated to the mobile sensors was considered in [27]. Such a containment policy aimed at providing limited local-in-space control action of the sensors that were within proximity of the moving source over the duration of a given time interval.

For the sake of exposition, the process under consideration here is governed by a one-dimensional diffusion-advection process, and the results are extendable to the two-dimensional case with minor adjustments for the sensor motion. Moreover, such a multidimensional extension requires attention to some technical issues pertaining to the existence and uniqueness of solutions to certain evolution equations with non–simply connected spatial domains, having nonsmooth boundaries, both internal and external, and the well-posedness of Lyapunov functions and their time derivatives that are subsequently used for the stability analysis. Additional conditions may also have to be imposed on the initial condition as well. However, the abstract framework is the same, as are the sensor navigation policies.

The paper is organized as follows. The diffusion process with its abstract framework formulation, along with the design of a state estimator, are summarized in section 2. The guidance policies (path planning) for the mobile sensors along with the requisite stability results are presented in section 3. Numerical studies of a one-dimensional diffusion-advection process utilizing both proposed guidance policies are reported in section 4, with conclusions and future research following in section 5.

2. Mathematical formulation and problem statement.

All notation in this paper is standard. For Banach spaces X and Y, $\mathcal{L}(X,Y)$ denotes the space of bounded linear operators from X into Y. All inner products $\langle \cdot,\cdot \rangle$ are assumed to be linear in their first argument and to be conjugate linear in the second. Additionally, $\langle \phi,\psi \rangle \triangleq \langle \phi,\psi \rangle_{X,Y}$ denotes the action of the linear functional $\psi \in Y$ on the element $\phi \in X$, and $\langle \psi,\phi \rangle \triangleq \langle \psi,\phi \rangle_{Y,X}$ denotes the actions of the conjugate linear functional $\psi \in Y$ on the element $\phi \in X$.

2.1. One-dimensional diffusion process. The diffusion process under consideration is modeled by a parabolic PDE on the bounded interval $\Omega = [0, \ell] \subset \mathbb{R}$. The state of the system is denoted by $x(t, \xi)$, where $\xi \in \Omega$ denotes the spatial variable and $t \in [0, \infty[$ is the time variable. The PDE is given by

$$\frac{\partial x}{\partial t}(t,\xi) = a_1 \frac{\partial^2 x}{\partial \xi^2}(t,\xi) - a_2 \frac{\partial x}{\partial \xi}(t,\xi) - a_3 x(t,\xi) + b_1(\xi)w(t) + b_2(\xi)u(t), \tag{2.1}$$

where $a_1, a_2, a_3 > 0$, along with Dirichlet boundary conditions $x(t, 0) = x(t, \ell) = 0$ and initial condition $x(0, \xi) = x_0(\xi) \in L_2(\Omega)$. The function $b_1(\xi) \in L_2(\Omega)$ denotes the spatial distribution of the process noise and $w(t)$ denotes its temporal component. Similarly, the function $b_2(\xi) \in L_2(\Omega)$ denotes the spatial distribution of the input function, while $u(t)$ denotes its temporal component. For the case of spatiotemporally moving inputs, or mobile controls, one may consider $b_2(\xi; \xi_a(t))$, where $\xi_a(t)$ denotes the time-varying location of the mobile actuating device. Such mobile actuators and their associated control policies were considered in [12, 23, 24, 25, 27, 30, 31, 32, 33, 34, 35, 53].

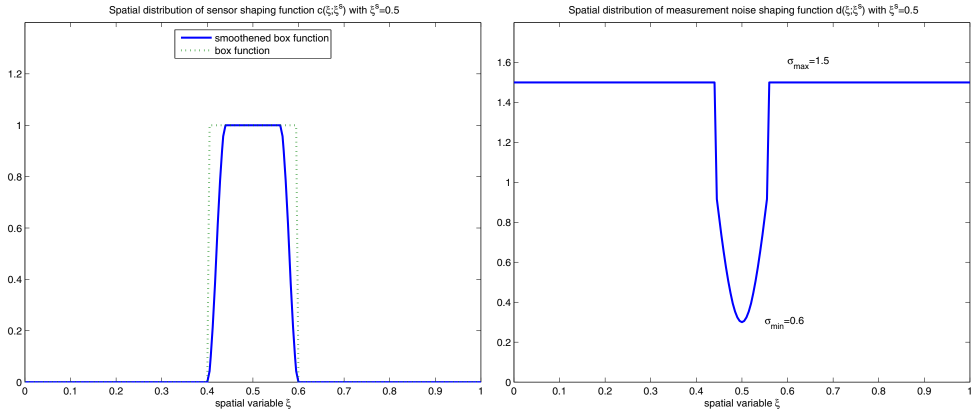
Spatially distributed measurements from m sensors are assumed to be available over the spatial intervals $[\xi_k^s - \Delta\xi \leq \xi \leq \xi_k^s + \Delta\xi]$, $k = 1, 2, \ldots, m$,

$$y(t,\xi;\xi^s) = \begin{bmatrix} c(\xi;\xi_1^s)x(t,\xi) + d(\xi;\xi_1^s)v(t) \\ c(\xi;\xi_2^s)x(t,\xi) + d(\xi;\xi_2^s)v(t) \\ \vdots \\ c(\xi;\xi_m^s)x(t,\xi) + d(\xi;\xi_m^s)v(t) \end{bmatrix},$$

where ξ_k^s denotes the kth sensor position within the domain $[0, \ell]$, $\xi^s = [\xi_1^s, \ldots, \xi_m^s] \in \mathbb{R}^m$ denotes the vector of sensor locations, $\Delta\xi$ denotes the one-half spatial support of the sensing device, and $c(\xi; \xi_k^s) \in L_2(0, \ell)$, $k = 1, 2, \ldots, m$, denotes the output shaping function associated with the kth sensor. The spatial distribution (shaping function) of the measurement noise is denoted by $d(\xi; \xi_k^s) \in L_2(0, \ell)$ and is similarly defined on the interval $[\xi_k^s - \Delta\xi \leq \xi \leq \xi_k^s + \Delta\xi]$. Its temporal component is denoted by $v(t)$. Example distributions for the output measurement and output noise shaping functions are depicted in Figure 2.1. Other distributions may also be considered. Examples include a Gaussian function or any other polynomial or trigonometric function. Figure 2.1(a) depicts both the box function and its smoothed approximation. The smoothed approximation is necessary for both regularity and numerical implementation requirements. Regarding regularity, smoothing guarantees well-posedness and certain system-theoretic properties, such as approximate observability, of the infinite dimensional system described below, to easily follow from already established results. On the numerical implementation side, smoothing aims to avoid Gibb's type phenomena in the numerical approximation of nonsmooth functions such as the box function. In fact, the following polynomial representation is used here for the smoothed approximation of the box function:

$$c(\xi;\xi_k^s) = \begin{cases} 1 & \text{if } \xi \in [\xi_k^s - 0.6\Delta\xi, \xi_k^s + 0.6\Delta\xi], \\ 1 - 3\xi_{l_k}^2 - 2\xi_{l_k}^3 & \text{if } \xi \in [\xi_k^s - \Delta\xi, \xi_k^s - 0.6\Delta\xi], \\ 1 - 3\xi_{r_k}^2 + 2\xi_{r_k}^3 & \text{if } \xi \in [\xi_k^s + 0.6\Delta\xi, \xi_k^s + \Delta\xi], \\ 0 & \text{otherwise}, \end{cases}$$

where $\xi_{r_k} = \frac{\xi - \xi_k^s - 0.6\Delta\xi}{0.4\Delta\xi}$ and $\xi_{l_k} = \frac{\xi - \xi_k^s + 0.6\Delta\xi}{0.4\Delta\xi}$. A similar cubic polynomial is used to smooth the spatial distribution of the measurement noise. As shown in Figure 2.1(b),



(a) $c(\xi;\xi^s)$ (b) $d(\xi;\xi^s)$

FIG. 2.1. *Spatial distributions* $c(\xi;\xi^s)$ *and* $d(\xi;\xi^s)$ *for* $\xi^s = \ell/2$.

the value of $d(\xi;\xi_k^s)$ is equal to σ_{max} outside the sensor range $[\xi_k^s - \Delta\xi,\ \xi_k^s + \Delta\xi]$, but its contribution to the output measurement is removed by multiplying it by the box function, thereby eliminating the introduction of noise outside the sensor range, i.e., excluding noise from the spatial interval $[0,\ell] \setminus [\xi_k^s - \Delta\xi,\ \xi_k^s + \Delta\xi]$. The noise effects are smaller at the center of the sensor range and increase as one moves away from the center. Similarly, the expression for $d(\xi;\xi^s)$ is

$$d(\xi;\xi_k^s) = \begin{cases} \sigma_{max} & \xi \in [\xi_k^s - \Delta\xi, \xi_k^s - \Delta\xi]^c, \\ \sigma_{max}(1 - 3\xi_{r_k}^2 + 2\xi_{r_k}^3) & \xi \in [\xi_k^s + 0.6\Delta\xi, \xi_k^s + \Delta\xi], \\ \sigma_{max}(1 - 3\xi_{l_k}^2 - 2\xi_{l_k}^3) & \xi \in [\xi_k^s - \Delta\xi, \xi_k^s - 0.6\Delta\xi], \\ \sigma_{min} + 2(\sigma_{max} - \sigma_{min})\Xi^2 - (\sigma_{max} - \sigma_{min})\Xi_k^4 & \text{otherwise}, \end{cases}$$

where $[\xi_k^s - \Delta\xi,\ \xi_k^s - \Delta\xi]^c = [0,\ell] \setminus [\xi_k^s - \Delta\xi,\ \xi_k^s + \Delta\xi]$ and $\Xi_k = \frac{\xi - \xi_k^s}{\Delta\xi}$.

When mobile sensors are considered, then one has

$$y(t;\xi^s) = \begin{bmatrix} c(\xi;\xi_1^s(t))x(t,\xi) + d(\xi;\xi_1^s(t))v_1(t) \\ c(\xi;\xi_2^s(t))x(t,\xi) + d(\xi;\xi_2^s(t))v_2(t) \\ \vdots \\ c(\xi;\xi_m^s(t))x(t,\xi) + d(\xi;\xi_m^s(t))v_m(t). \end{bmatrix}, \tag{2.2}$$

which explicitly models the sensor position motion via the time-variation of the second argument of the output measurement shaping function $c(\xi;\xi_i^s(t))$ and noise measurement shaping function $d(\xi;\xi_i^s(t))$ for the ith sensing device.

2.2. Abstract formulation. For well-posedness and stability of the proposed estimation scheme, the above PDE given in (2.1), (2.2) will be viewed in an abstract framework. Such an abstract framework includes a larger class of PDEs, and hence the results can be easily applied to any member of this class of systems.

We let $\mathcal{X}$ be a Hilbert space with inner product $\langle\cdot,\cdot\rangle$ and corresponding induced norm $|\cdot|$. Let $\mathcal{V}$ be a reflexive Banach space with norm denoted by $\|\cdot\|$, and assume that $\mathcal{V}$ is embedded densely and continuously in $\mathcal{X}$ [77, 90]. Let $\mathcal{V}^*$ denote the conjugate dual of $\mathcal{V}$ (in other words, the space of continuous conjugate linear functionals on $\mathcal{V}$) and $\|\cdot\|_*$ denote the usual uniform operator norm on $\mathcal{V}^*$. It follows that

$$\mathcal{V} \hookrightarrow \mathcal{X} \hookrightarrow \mathcal{V}^*, \tag{2.3}$$

with both embeddings dense and continuous [22, 80]. Specifically, we assume that

$$|\phi| \leq c\|\phi\|, \qquad \phi \in \mathcal{V}, \tag{2.4}$$

for some positive constant c. The notation $\langle\cdot,\cdot\rangle$ will also be used to denote the duality pairing between $\mathcal{V}^*$ and $\mathcal{V}$ induced by the continuous and dense embeddings given in (2.3); that is, for $\phi \in \mathcal{V}^*$ and $\psi \in \mathcal{V}$, $\langle\phi,\psi\rangle$ denotes the action of the bounded linear functional ϕ on the vector ψ. This quantity simply reduces to $\langle\phi,\psi\rangle$ if $\phi \in \mathcal{X}$; i.e., the value of ϕ acting on ψ is equal to the $\mathcal{X}$ inner product of ϕ and ψ.

We consider a linear operator $\mathcal{A} : \mathcal{V} \to \mathcal{V}^*$ satisfying the following assumptions:

(A1) $\mathcal{V} \to \mathcal{V}^*$-*boundedness*: There exists $\alpha > 0$ such that

$$|\langle\mathcal{A}\phi,\psi\rangle| \leq \alpha\|\phi\|\,\|\psi\| \qquad \text{for } \phi,\psi \in \mathcal{V}.$$

(A2) $\mathcal{V}$-*coercivity*: the operator $-\mathcal{A}$ is coercive, i.e.,

$$\text{Re}\,\langle-\mathcal{A}\phi,\phi\rangle \geq \beta\|\phi\|^2 \qquad \text{for some positive } \beta \text{ and } \phi \in \mathcal{V}.$$

Additionally, we may impose the following symmetry condition which, while it simplifies the stability analysis, nonetheless restricts the class of systems (i.e., diffusion processes) to which the proposed sensor navigation and state estimation policy is applicable.

(A3) *Symmetry*: the operator $\mathcal{A}$ is symmetric:

$$\langle\mathcal{A}\phi,\psi\rangle = \overline{\langle\mathcal{A}\phi,\psi\rangle} \qquad \text{for all } \phi,\psi \in \mathcal{V}.$$

For ease of exposition, we have assumed that the operator $\mathcal{A}$ is time invariant. However, it is relatively straightforward to extend all of the results in this paper to the case of a time-dependent operator $\mathcal{A}(t)$, $t \geq 0$. One need only make some standard assumptions on the regularity of the map $t \to \mathcal{A}(t)$, $t \geq 0$, for the present results to remain valid [8, 61, 76, 85].

We consider the disturbance operator $\mathcal{B}_1 : \mathbb{R} \to \mathcal{V}^*$ and the *input* operator $\mathcal{B}_2 : \mathbb{R} \to \mathcal{V}^*$. When the control and disturbance signals are assumed to be square integrable, i.e., yielding $\mathcal{B}_1 w + \mathcal{B}_2 u \in L_2(0,t,\mathcal{V}^*)$, and $x(0) = x_0 \in \mathcal{X}$, then the initial value problem (IVP)

$$\frac{d}{dt}x(t) = \mathcal{A}x(t) + \mathcal{B}_1 w(t) + \mathcal{B}_2 u(t), \qquad x_0 \in \mathcal{X}, \tag{2.5}$$

is well-posed. By a solution to the above (IVP), we mean a weak solution [76]; this means a function $x \in L_2(0,t;\mathcal{V})$ with $\frac{d}{dt}x \in L_2(0,t;\mathcal{V}^*)$ for all $t > 0$ that satisfies (2.5) [76, 90].

Following [21], the PDE in (2.1) may be expressed in the abstract form (2.5). The state space in this case is $\mathcal{X} = L_2(0,\ell)$, where $x(t,\cdot) = \{x(t,\xi),\, 0 \leq \xi \leq \ell\}$ denotes the state. The space $\mathcal{V}$ is identified by the Sobolev space $\mathcal{V} = H_0^1(0,\ell)$. In the remainder of the paper we will, with a slight abuse of notation, use $x(t)$ as the solution to the evolution equation (2.5) and use $x(t,\xi)$ as the solution to the PDE (2.1). Under the above representation, the system's second order (strongly) elliptic operator $\mathcal{A}$ and its domains are given by [22]:

$$\begin{aligned} \mathcal{A}\phi &= a_1\tfrac{d^2\phi}{d\xi^2} - a_2\tfrac{d\phi}{d\xi} - a_3\phi, \qquad \phi \in \text{Dom }(\mathcal{A}),\\ \text{Dom }(\mathcal{A}) &= H^2(0,\ell)\cap H_0^1(0,\ell)\\ &= \{\psi \in L_2(0,\ell)\,|\,\psi,\psi^{'} \text{ abs. continuous and } \psi(0)=0=\psi(\ell)\}. \end{aligned}$$

We now verify the boundedness and coercivity assumptions (A1) and (A2) for the above system:

$$\begin{aligned} |\langle \mathcal{A}\phi,\psi\rangle| &= \left|\int_0^\ell \left(a_1\frac{d^2\phi(\xi)}{d\xi^2} - a_2\frac{d\phi(\xi)}{d\xi} - a_3\phi(\xi)\right)\psi(\xi)\,d\xi\right|\\ &\le a_1\left|\int_0^\ell \frac{d^2\phi(\xi)}{d\xi^2}\psi(\xi)\,d\xi\right| + a_2\left|\int_0^\ell \frac{d\phi(\xi)}{d\xi}\psi(\xi)\,d\xi\right| + a_3\left|\int_0^\ell \phi(\xi)\psi(\xi)\,d\xi\right|\\ &\le a_1\sqrt{\int_0^\ell\left(\frac{d\phi(\xi)}{d\xi}\right)^2 d\xi}\sqrt{\int_0^\ell\left(\frac{d\psi(\xi)}{d\xi}\right)^2 d\xi}\\ &\quad +a_2\sqrt{\int_0^\ell\left(\frac{d\phi(\xi)}{d\xi}\right)^2 d\xi}\sqrt{\int_0^\ell \psi^2(\xi)\,d\xi} + a_3\sqrt{\int_0^\ell \phi^2(\xi)\,d\xi}\sqrt{\int_0^\ell \psi^2(\xi)\,d\xi}\\ &= a_1\|\phi\|\,\|\psi\| + a_2\|\phi\|\,|\psi| + a_3|\phi|\,|\psi|\\ &\le a_1\|\phi\|\,\|\psi\| + a_2\|\phi\|c\|\psi\| + a_3c^2\|\phi\|\,\|\psi\| = \left(a_1 + a_2c + a_3c^2\right)\|\phi\|\,\|\psi\|, \end{aligned}$$

where we used the triangle inequality in the first step and used the fact that the space $\mathcal{V}$ is embedded in $\mathcal{X}$. This proves boundedness. To show coercivity, note that

$$\begin{aligned} \langle -\mathcal{A}\phi,\phi\rangle &= \int_0^\ell -\left(a_1\frac{d^2\phi(\xi)}{d\xi^2} - a_2\frac{d\phi(\xi)}{d\xi} - a_3\phi(\xi)\right)\phi(\xi)\,d\xi\\ &= -a_1\int_0^\ell \frac{d^2\phi(\xi)}{d\xi^2}\phi(\xi)\,d\xi + a_2\int_0^\ell \frac{d\phi(\xi)}{d\xi}\phi(\xi)\,d\xi + a_3\int_0^\ell \phi(\xi)\phi(\xi)\,d\xi\\ &\ge a_1\int_0^\ell\left(\frac{d\phi(\xi)}{d\xi}\right)^2 d\xi + a_2\int_0^\ell \frac{d}{d\xi}\phi^2(\xi)\,d\xi + a_3\int_0^\ell \phi^2(\xi)\,d\xi\\ &= a_1\|\phi\|^2 + a_2|\phi^{'}|^2 + a_3|\phi|^2 \ge a_1\|\phi\|^2. \end{aligned}$$

It should be noted, however, that due to the presence of a nonzero coefficient a_2, the operator is not symmetric.

The input operator is given by

$$\mathcal{B}_2u(t) = b_2(\xi)u(t), \qquad \mathcal{B}_2 \in \mathcal{L}(\mathbb{R},\mathcal{X}).$$

The disturbance (process noise) operator $\mathcal{B}_1$ is given similarly by

$$\mathcal{B}_1w(t) = b_1(\xi)w(t), \qquad \mathcal{B}_1 \in \mathcal{L}(\mathbb{R},\mathcal{X}).$$

Similarly, the output equation (2.2) may be written as

$$y(t;\xi^s) = \mathcal{C}(\xi^s(t))x(t) + \mathcal{D}(\xi^s(t))v(t), \tag{2.6}$$

where the output measurement and noise operators are parameterized by the sensor location vector ξ^s. These operators are given via $\mathcal{C}(\cdot) : \mathcal{V} \rightarrow \underbrace{\mathcal{V}^* \times \mathcal{V}^* \times \cdots \times \mathcal{V}^*}_{m}$ by

$$\langle \mathcal{C}(\xi^s(t))\phi, \psi \rangle = \begin{bmatrix} \displaystyle\int_0^\ell c(\xi; \xi_1^s(t))\phi(\xi)\psi(\xi)\, d\xi \\ \displaystyle\int_0^\ell c(\xi; \xi_2^s(t))\phi(\xi)\psi(\xi)\, d\xi \\ \vdots \\ \displaystyle\int_0^\ell c(\xi; \xi_m^s(t))\phi(\xi)\psi(\xi)\, d\xi \end{bmatrix},$$

and via $\mathcal{D}(\cdot) : \mathbb{R}^m \rightarrow \underbrace{\mathcal{V}^* \times \mathcal{V}^* \times \cdots \times \mathcal{V}^*}_{m}$ by

$$\langle \mathcal{D}(\xi^s(t))\nu, \psi \rangle = \begin{bmatrix} \displaystyle\int_0^\ell d(\xi; \xi_1^s(t))\psi(\xi)\, d\xi\, \nu_1 \\ \displaystyle\int_0^\ell d(\xi; \xi_2^s(t))\psi(\xi)\, d\xi\, \nu_2 \\ \vdots \\ \displaystyle\int_0^\ell d(\xi; \xi_m^s(t))\psi(\xi)\, d\xi\, \nu_m \end{bmatrix}.$$

2.3. Problem statement. The problem at hand is to propose a state estimator for the evolution system (2.5), with measurements given by (2.2), and to provide a motion planning strategy of the mobile sensors in (2.2) in order to yield a more efficient state estimator.

2.4. State estimation process with time-varying output operator. For an arbitrary but fixed sensor location ξ^s, one may consider the associated state estimator in $\mathcal{X}$,

$$\dot{\widehat{x}}(t) = \mathcal{A}\widehat{x}(t) + \mathcal{B}_2 u(t) + \mathcal{L}(\xi^s)\Big(y(t) - \mathcal{C}(\xi^s)\widehat{x}(t)\Big), \tag{2.7}$$

where $\widehat{x}(0) = \widehat{x}_0 \in \mathcal{X}$ with $\widehat{x}(0) \neq x(0)$, and $\mathcal{L}(\xi^s) : \underbrace{\mathcal{V}^* \times \mathcal{V}^* \times \cdots \times \mathcal{V}^*}_{m} \rightarrow \mathcal{V}$ is the associated ξ^s-parameterized observer gain derived from either a Kalman or a Luenberger filter design.

The *state estimation error* $e(t) \triangleq x(t) - \widehat{x}(t)$ for (2.5) is governed by the following evolution equation:

(2.8)
$$\begin{aligned} \dot{e}(t) &= \mathcal{A}e(t) - \mathcal{L}(\xi^s)\left(y(t) - \mathcal{C}(\xi^s)\widehat{x}(t)\right) + \mathcal{B}_1 w(t), \\ &= \left(\mathcal{A} - \mathcal{L}(\xi^s)\mathcal{C}(\xi^s)\right) e(t) + \mathcal{B}_1 w(t) - \mathcal{L}(\xi^s)\mathcal{D}(\xi^s)v(t), \end{aligned} \qquad e(0) = x(0) - \widehat{x}(0) \in \mathcal{X}.$$

The associated distributed *output estimation error* $\varepsilon(t; \xi^s)$ corresponding to m sensor locations represented by the vector ξ^s is given by

$$\varepsilon(t; \xi^s) \triangleq y(t; \xi^s) - \mathcal{C}(\xi^s)\widehat{x}(t). \tag{2.9}$$

Next, this output error will be used to generate the navigation policies for the mobile sensors.

3. Navigation of spatially mobile sensors. The sensor locations ξ^s considered above are now allowed to vary with time, and thus the above observation and measurement noise operators are time dependent. The associated state estimator is now given by (cf. (2.7))

$$\dot{\widehat{x}}(t) = \Big(\mathcal{A} - \mathcal{L}(\xi^s(t))\mathcal{C}(\xi^s(t))\Big)\widehat{x}(t) + \mathcal{B}_2 u(t) + \mathcal{L}(\xi^s(t))y(t), \tag{3.1}$$

which results in the estimation error equation (cf. (2.8))

$$\begin{aligned} \dot{e}(t) &= \Big(\mathcal{A} - \mathcal{L}(\xi^s(t))\mathcal{C}(\xi^s(t))\Big)e(t) + \mathcal{B}_1 w(t) - \mathcal{L}(\xi^s(t))\mathcal{D}(\xi^s(t))v(t), \\ e(0) &= x(0) - \widehat{x}(0) \in \mathcal{X}. \end{aligned} \tag{3.2}$$

Similarly, the output estimation error (cf. (2.9)) is given by

$$\varepsilon(t;\xi^s(t)) = \mathcal{C}(\xi^s(t))e(t) + \mathcal{D}(\xi^s(t))v(t). \tag{3.3}$$

One may consider an optimal sensor scheduling, as was developed in [7], to derive the position of the mobile sensors. While the resulting sensor guidance will be optimal, it would nonetheless result in computationally intensive implementation requiring the solution to differential Riccati operator equations. Motivated by computational considerations, we consider sensor guidance schemes that may forsake optimality for ease of implementation and reduction of the computational load. Additionally, we do not necessarily consider a finite horizon problem, and thus one may have to address the issue of observability. Thus, we assume that the sensor guidance scheme navigates the mobile sensors only in the spatial locations that render the system approximately observable [21]. To avoid such locations, we define the set of admissible locations as

$$\Theta_{adm} = \Big\{\xi_i^s \in \Omega \;\;:\;\; (\mathcal{C}(\xi^s), \mathcal{A}) \text{ is approximately observable}\Big\}. \tag{3.4}$$

Any sensor scheduling will then be constrained to the set Θ_{adm}. While we will not explicitly impose this condition, one may incorporate such an admissibility condition into a collision avoidance navigation scheme, whereby both undesirable locations and locations that render the system unobservable will be avoided.

The above error provides distributed information of the estimation error throughout the support of a given sensing device, i.e., over $[\xi_i^s - \Delta\xi,\, \xi_i^s + \Delta\xi]$. Using only this spatially distributed error, we propose two guidance policies. The first guidance policy moves the center of the ith sensing device so as to minimize its spatial L_∞ norm within the spatial interval $[\xi_i^s - \Delta\xi,\, \xi_i^s + \Delta\xi]$. Such a guidance policy renders the resulting infinite dimensional system a switched system whereby the position of the m sensors changes at discrete time instances. The second guidance policy considers the global distributed error associated with the nominal noise-free process, and by embedding the sensor position in the process dynamics, a guidance law is derived using Lyapunov stability arguments.

3.1. Case 1: Guidance using localized measurement error.

Assuming that a given sensor can only move a maximum distance of $\pm\Delta\xi$ from its current position $\xi_i^s(t_k)$, i.e., move anywhere within $[\xi_i^s(t_k) - \Delta\xi,\, \xi_i^s(t_k) + \Delta\xi]$, and taking into account velocity constraints which translate into restrictions on the frequency of the switching positions, we consider the time instances $t_0 + k\Delta t$, $k = 0, 1, 2, \ldots$. The proposed sensor position switching is given by

$$\xi_i^s(t_{k+1}) = \arg \max_{\xi_i^s(t_k) - \Delta\xi \le \xi \le \xi_i^s(t_k) + \Delta\xi} |\varepsilon(t_k, \xi; \xi_i^s(t_k))| \tag{3.5}$$

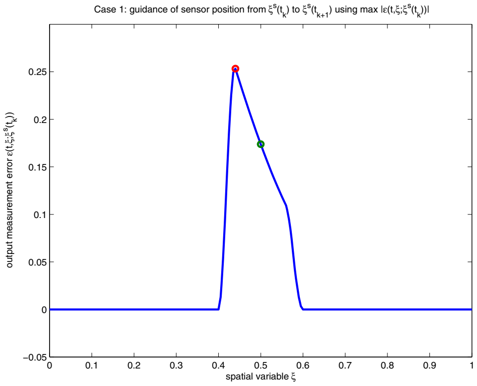


Fig. 3.1. *Guidance of moving sensor from position $\xi^s(t_k)$ (lower dot) to position $\xi^s(t_{k+1})$ (upper dot) using localized measurement error $\varepsilon(t,\xi;\xi_i^s(t_k))$.*

for $i = 1, 2, \ldots, m$, and which basically finds the maximum of the spatially distributed measurement error over the domain of definition $[\xi_i^s(t_k) - \Delta\xi,\ \xi_i^s(t_k) + \Delta\xi]$ of the current sensor position $\xi_i^s(t_k)$, and moves the ith sensor to that maximum. Figure 3.1 depicts a scenario with the current $\xi^s(t_k)$ and subsequent sensor position $\xi^s(t_{k+1})$.

Remark 3.1. *Note that if $\varepsilon(t,\xi;\xi_i^s(t_k)) = 0$ inside the ith sensor domain, then sensor i will not move. If $\varepsilon(t,\xi;\xi_i^s(t_k)) = 0$ for all sensors i, then none of the sensors will move, even if the error is nonzero outside the sensor domains. In the spatially decoupled case (i.e., no PDE is satisfied), this is problematic since the error outside all sensor domains is not "transmitted" to the error within the sensor domains. In this case, the sensors are immobile with a nonzero global error. This situation was called Condition* **C1** *in* [51]. *To resolve this issue, the authors in* [51] *propose perturbation control laws that transfer sensors with zero error within the domain to a neighborhood of a point (outside the sensor domain) with nonzero error. Once there, the control law is switched back to* (3.5). *It was shown in* [51] *that such a switching policy causes the coverage error to converge to zero and that infinite switching is impossible if the domain $\mathcal{D}$ is compact.*

For the spatially coupled case, under the assumption that the PDE system is approximately observable, which is guaranteed by restricting the sensor motion to the set Θ_{adm}, one has that the error inside the sensor domains is zero if and only if the global error is zero. Hence Condition **C1** *(that error inside the sensor domain is zero with nonzero global error) described above will never occur and switching is not required.*

For the specific case of the observer operator gain $\mathcal{L}(\xi^s(t)) = \mathcal{C}^*(\xi^s(t))$, the above state error (2.8) reduces to the switched infinite dimensional system

$$\begin{aligned} \dot{e}(t) &= \Big(\mathcal{A} - \mathcal{C}^*(\xi^s(t))\mathcal{C}(\xi^s(t))\Big)e(t) + \mathcal{B}_1 w(t) - \mathcal{C}^*(\xi^s(t))\mathcal{D}(\xi^s(t))v(t), \\ e(0) &= x(0) - \widehat{x}(0) = e_0 \in \mathcal{X}. \end{aligned} \tag{3.6}$$

We examine the stability of the above switched system (3.5), (3.6) within the context of switched infinite dimensional systems. We use the notation Ξ_k to denote the sensor position throughout the time interval $[t_k, t_k + \Delta t)$; i.e., the sensors will maintain the same position $\xi^s(t)$ for the duration of the time interval $[t_k, t_k + \Delta t)$,

$$\xi^s(t) = \Xi^s, \qquad t \in [t_k, t_k + \Delta t).$$

We define the operators $\mathcal{A}_i = \mathcal{A} - \mathcal{C}^*(\Xi_i)\mathcal{C}(\Xi_i)$ for $\Xi_i \in \Theta_{adm}$. In view of the above, we consider the family of infinitesimal generators $\mathfrak{A} = \{\mathcal{A}_i,\ i \in \mathcal{I}\}$ on $\mathcal{X}$ parameterized by some index set $\mathcal{I}$. We let $\sigma : [0,\infty) \to \mathcal{I}$ be a piecewise constant function of time, termed the *switching signal.* Additionally, we define the operators $\mathfrak{B}_i : \mathcal{U} = \mathbb{R} \oplus \mathbb{R}^m \to \mathcal{X}$ and $\mathfrak{C}_i : \mathcal{X} \oplus \mathbb{R}^m \to \mathcal{Y}$,

$$\begin{aligned} \mathfrak{B}_i v(t) &= \mathcal{B}_1 w(t) - \mathcal{C}^*(\xi_i^s)\mathcal{D}(\xi_i^s)v(t), \qquad && v \in \mathcal{U}, \\ \mathfrak{C}_i \chi(t) &= \mathcal{C}(\xi_i^s)e(t) + \mathcal{D}(\xi_i^s)v(t), && \chi \in \mathcal{X} \oplus \mathbb{R}^m. \end{aligned}$$

Let $(S_p)_{p\in\mathcal{I}}$, for some index set $\mathcal{I}$, be a family of linear continuous time systems which, for each fixed $p \in \mathcal{I}$, is given by a state linear system $(\mathcal{A}_p, \mathfrak{B}_p, \mathfrak{C}_p)$,

$$(S_p) : \begin{cases} \dot{e}(t) = \mathcal{A}_p e(t) + \mathfrak{B}_p v(t), \\ \epsilon(t) = \mathfrak{C}_p \chi(t), \end{cases} \tag{3.7}$$

where the operator $\mathcal{A}_p = \mathcal{A} - \mathcal{C}^*(\Xi_p)\mathcal{C}(\Xi_p)$ is the infinitesimal generator of an exponentially stable semigroup $T_p(t)$ on the Hilbert space $\mathcal{X}$ for all $\Xi_p \in \Theta_{adm}$. For each $p \in \mathcal{I}$, the operators $\mathfrak{B}_p$ and $\mathfrak{C}_p$ are bounded linear operators from a Hilbert space $\mathcal{U}$ to $\mathcal{X}$ and from $\mathcal{X} \oplus \mathbb{R}^m$ to a Hilbert space $\mathcal{Y}$, respectively. To the family $(S_p)_{p\in\mathcal{I}}$, we associate the set

$$\Sigma = \{\sigma \mid \sigma : [t_0,\infty) \to \mathcal{I} \text{ piecewise constant}\}$$

of all possible switches between the given systems. The *family of switched systems* $((S_p)_{p\in\mathcal{I}}, \Sigma)$ taken under consideration are the hybrid dynamical systems consisting of the family of continuous time systems $(S_p)_{p\in\mathcal{I}}$ together with all switching rules $\sigma \in \Sigma$, all initial states $e(0) = e_0 \in \mathcal{X}$, and all inputs $v \in L_2([t_0,\infty);\mathcal{U})$. For each given switching function σ, denote the finite set of switching time instants associated to σ by $t_0 < t_1 < t_2 < \cdots$, where $k(\sigma) \in \mathbb{N} \setminus \{0\}$. Here, $k(\sigma) - 1$ denotes the number of discontinuities for the piecewise continuous function σ. For $k(\sigma) = 1$, the no-switching case is obtained. Therefore we impose $k(\sigma) \geq 2$ as a necessary condition for the nontriviality of the problem; i.e., there exists *at least* one switch.

By a solution, we mean that given a switching function σ, an initial condition e_0 and an input v, then, on each interval $\{[t_i, t_{i+1}]\}_{i=0}^{k(\sigma)-1}$, the state $e_\sigma(t)$ of the switched system $((S_p)_{p\in\mathcal{P}}, \sigma)$ is the mild solution of the Cauchy problem (3.7) [21], i.e. for $t_i \leq t \leq t_{i+1}$,

$$\begin{aligned} e_\sigma^i(t) &= T_\sigma(t)e_\sigma(t_i) + \int_{t_i}^t T_{\sigma(t_i)}(t-s)\mathfrak{B}_{\sigma(t_i)}v(s)ds, \\ \epsilon_\sigma^i(t) &= \mathcal{C}_{\sigma(t_i)}e_\sigma^i(t) + \mathcal{D}_{\sigma(t_i)}\nu(t). \end{aligned} \tag{3.8}$$

We make the assumption that the resulting hybrid system is not a jump system.

ASSUMPTION 3.1. *The initial conditions for the error state at the beginning of each interval* $\{[t_i, t_{i+1}]\}_{i=1}^{k(\sigma)-1}$ *are given by* $\{e_\sigma(t_i)\}_{i=1}^{k(\sigma)-1}$, *and they are considered to be the end values of the solution on the preceding time interval, i.e.,*

$$\overbrace{e_\sigma^{i+1}(t_{i+1})}^{initial\ value\ at\ [t_{i+1},t_{i+2}]} = \overbrace{e_\sigma^{i}(t_{i+1})}^{final\ value\ at\ [t_i,t_{i+1}]}.$$

Based on the above assumptions, the following then leads to the existence of solutions of the error system (3.6). We state only the result since the proof follows in a similar fashion to the one presented in [53] for scheduled actuators. As a consequence of the well-posedness, which along with the square integrability of the input signals and the exponential stability of the semigroups associated with each sensor position within Θ_{adm}, one also has convergence of the state estimation error e to zero.

THEOREM 3.1. *Assume that the operator $\mathcal{A}$ satisfies the coercivity and boundedness conditions* (A1) *and* (A2). *Furthermore, assume that for each $\xi^s \in \Theta_{adm}$, the operator $\mathcal{A} - \mathcal{C}^*(\xi^s)\mathcal{C}(\xi^s)$ generates an exponentially stable semigroup on $\mathcal{X}$ and that the control input u, along with the process and measurement noise, is square integrable, in the sense of $\mathcal{B}_1 w + \mathcal{B}_2 u \in L_2(0,\infty;\mathcal{X})$ and $\mathcal{B}_1 w - \mathcal{C}^*(\xi^s)\mathcal{D}(\xi^s)\nu \in L_2(0,\infty;\mathcal{X})$. Then the state error system* (3.6)*, along with the switching policy* (3.5)*, is well posed. As a consequence of that, the state estimator* (3.1)*, with the switching policy* (3.5)*, is also well posed.*

REMARK 3.2. *It should be noted that the fact that the operator in the evolution equation* (3.7) *is the infinitesimal generator of an exponentially stable semigroup (uniformly for each $\Xi_p \in \Theta_{adm}$), along with the fact that the forcing term $\mathfrak{B}_p \upsilon$ is square integrable, immediately yields exponential stability of* (3.7) *for the nonswitched case. For the switched case with the guidance policy given by the switching rule for the sensor location in* (3.5) *and excluding jump systems using Assumption* 3.1*, one can similarly show convergence of the state estimation error to zero using similar arguments that were used in the stability of diffusion systems with scheduled actuators in* [53].

3.2. Case 2: Guidance using global estimation error. Since the estimation error is available only at the spatial support of the sensors, we consider the idealized process, given by

$$\begin{aligned} \dot{\overline{x}}(t) &= \mathcal{A}\overline{x}(t) + \mathcal{B}_2 u(t), \\ \overline{y}(t) &= \mathcal{C}(\xi^s(t))\overline{x}(t), \end{aligned} \tag{3.9}$$

and define the *nominal estimation error* $\overline{e}(t) \triangleq \overline{x}(t) - \widehat{x}(t)$ governed by

$$\begin{aligned} \dot{\overline{e}}(t) &= \Big(\mathcal{A} - \mathcal{L}(\xi^s(t))\mathcal{C}(\xi^s(t))\Big)\overline{e}(t) + \mathcal{L}(\xi^s(t))\,(\overline{y} - y) \\ &= \mathcal{A}_o(\xi^s(t))\overline{e}(t) + \mathcal{L}(\xi^s(t))\Big(\mathcal{C}(\xi^s(t))\,(\overline{x} - x) - \mathcal{D}(\xi^s(t))v(t)\Big), \end{aligned} \tag{3.10}$$

with $\overline{e} \in \mathcal{X}$, and where $\mathcal{A}_o(\xi^s(t)) \triangleq \mathcal{A} - \mathcal{L}(\xi^s(t))\mathcal{C}(\xi^s(t))$. The above error is generated online and simulates the idealized process (2.5), (2.6) in the absence of process and measurement noise and possibly exogenous/disturbance inputs. In a similar fashion as in the case of the process operator $\mathcal{A}$, we make similar boundedness and coercivity assumptions for $\mathcal{A}_o$ in (3.10) with the constants α, β now replaced by α_o, β_o.

We consider the weighted $L_2(0,\ell)$ inner product

$$\langle \overline{e}, \overline{e} \rangle_g \triangleq \langle \overline{e}(t), g\overline{e}(t) \rangle, \tag{3.11}$$

where the normalized weighting function $g(\xi) > 0$ for all $\xi \in \Omega$, $g \in \mathcal{L}_\infty(\Omega)$, is also known as the *distribution density function.* This function may be used to emphasize the need to cover some intervals in Ω more than others. Similar to the notation for the boundedness and coercivity constants, when the weighted inner product is used, those constants will include g as a second subscript. For the specific PDE under consideration, the weighted inner product becomes

$$\langle \overline{e}, \overline{e} \rangle_g = \langle \overline{e}(t), \overline{e}(t) \rangle_{L_2,g} = \int_0^\ell \overline{e}^2(t,\xi) g(\xi) d\xi,$$

To obtain the sensor guidance using Lyapunov stability–based arguments, we consider the following Lyapunov-like functional:

$$
\begin{aligned}
V(t;\overline{e},\xi^s(t)) &= -\frac{1}{2}\Big(\langle \overline{e}(t),\mathcal{A}_o(\xi^s)\overline{e}(t)\rangle_g + \langle \mathcal{A}_o(\xi^s)\overline{e}(t),\overline{e}(t)\rangle_g\Big) \\
&= -\left\langle \overline{e}(t), \left(\frac{\mathcal{A}_o(\xi^s)+\mathcal{A}_o^*(\xi^s)}{2}\right)\overline{e}(t)\right\rangle_g .
\end{aligned}
\tag{3.12}
$$

This function is a modified version of that used in [49] and has the following explanation. The function V represents the negative of the derivative of the weighted error norm (3.11) along the trajectories of the nominal estimation error (3.10). The reason we chose $-\frac{d}{dt}\|\overline{e}\|^2$ instead of $\|\overline{e}\|^2$ itself for the Lyapunov-like function is that if the latter is chosen, the expression for $\dot{V}$ will not involve the control variable $\dot{\xi}^s$.

For brevity, we suppress the dependence of the operator $\mathcal{A}_o(\xi^s)$ on the sensor position and we simply write $\mathcal{A}_o$. Since $\mathcal{A}_o$ is assumed coercive, then we see that the function V is positive for all nonzero $\bar{e}$ and is zero if and only if $\bar{e}$ is zero. The derivative of V along the trajectories of the nominal estimation error (3.10) is then given by

$$
\begin{aligned}
\frac{d}{dt}V &= -\Bigg\{ \langle \dot{\overline{e}},\mathcal{A}_o\overline{e}\rangle_{L_2,g} + \langle \overline{e},\mathcal{A}_o\dot{\overline{e}}\rangle_{L_2,g} + \left\langle \overline{e},\dot{\xi}^s\frac{\partial\mathcal{A}_o}{\partial\xi^s}\overline{e}\right\rangle_{L_2,g} \\
&\qquad + \langle \mathcal{A}_o\dot{\overline{e}},\overline{e}\rangle_{L_2,g} + \left\langle \dot{\xi}^s\frac{\partial\mathcal{A}_o}{\partial\xi^s}\overline{e},\overline{e}\right\rangle_{L_2,g} + \langle \mathcal{A}_o\overline{e},\dot{\overline{e}}\rangle_{L_2,g} \Bigg\} \\
&= -\Bigg\{ \langle \mathcal{A}_o\overline{e},\mathcal{A}_o\overline{e}\rangle_{L_2,g} + \langle \overline{e},\mathcal{A}_o(\mathcal{A}_o\overline{e})\rangle_{L_2,g} \\
&\qquad + \left\langle \overline{e},\dot{\xi}^s\frac{\partial(\mathcal{A}-\mathcal{L}(\xi^s)\mathcal{C}(\xi^s))}{\partial\xi^s}\overline{e}\right\rangle_{L_2,g} + \langle \mathcal{A}_o(\mathcal{A}_o\overline{e}),\overline{e}\rangle_{L_2,g} \\
&\qquad + \left\langle \dot{\xi}^s\frac{\partial(\mathcal{A}-\mathcal{L}(\xi^s)\mathcal{C}(\xi^s))}{\partial\xi^s}\overline{e},\overline{e}\right\rangle_{L_2,g} + \langle \mathcal{A}_o\overline{e},\mathcal{A}_o\overline{e}\rangle_{L_2,g} \\
&\qquad + \langle \mathcal{L}(\xi^s)\left(\mathcal{C}(\xi^s)(\overline{x}-x)-\mathcal{D}v\right),\mathcal{A}_o\overline{e}\rangle_{L_2,g} \\
&\qquad + \langle \overline{e},\mathcal{A}_o\mathcal{L}(\xi^s)\left(\mathcal{C}(\xi^s)(\overline{x}-x)-\mathcal{D}v\right)\rangle_{L_2,g} \\
&\qquad + \langle \mathcal{A}_o\mathcal{L}(\xi^s)\left(\mathcal{C}(\xi^s)(\overline{x}-x)-\mathcal{D}v\right),\overline{e}\rangle_{L_2,g} \\
&\qquad + \langle \mathcal{A}_o\overline{e},\mathcal{L}(\xi^s)\left(\mathcal{C}(\xi^s)(\overline{x}-x)-\mathcal{D}v\right)\rangle_{L_2,g} \Bigg\} \\
&= -\Bigg\{ \|\mathcal{A}_o\overline{e}\|^2_{L_2,g} + \langle \overline{e},\mathcal{A}_o(\mathcal{A}_o\overline{e})\rangle_{L_2,g} - \left\langle \overline{e},\dot{\xi}^s\frac{\partial(\mathcal{L}(\xi^s)\mathcal{C}(\xi^s))}{\partial\xi^s}\overline{e}\right\rangle_{L_2,g} \\
&\qquad + \langle \mathcal{A}_o(\mathcal{A}_o\overline{e}),\overline{e}\rangle_{L_2,g} - \left\langle \dot{\xi}^s\frac{\partial(\mathcal{L}(\xi^s)\mathcal{C}(\xi^s))}{\partial\xi^s}\overline{e},\overline{e}\right\rangle_{L_2,g} + \|\mathcal{A}_o\overline{e}\|^2_{L_2,g} \\
&\qquad + \langle \mathcal{L}(\xi^s)\left(\mathcal{C}(\xi^s)(\overline{x}-x)-\mathcal{D}v\right),\mathcal{A}_o\overline{e}\rangle_{L_2,g} \\
&\qquad + \langle \overline{e},\mathcal{A}_o\mathcal{L}(\xi^s)\left(\mathcal{C}(\xi^s)(\overline{x}-x)-\mathcal{D}v\right)\rangle_{L_2,g} \\
&\qquad + \langle \mathcal{A}_o\mathcal{L}(\xi^s)\left(\mathcal{C}(\xi^s)(\overline{x}-x)-\mathcal{D}v\right),\overline{e}\rangle_{L_2,g} \\
&\qquad + \langle \mathcal{A}_o\overline{e},\mathcal{L}(\xi^s)\left(\mathcal{C}(\xi^s)(\overline{x}-x)-\mathcal{D}v\right)\rangle_{L_2,g} \Bigg\}.
\end{aligned}
$$

We will examine each of the terms above separately: the second and fourth terms which, due to the symmetry of the weighted inner product, are identical and are given by

$$\langle \overline{e}, \mathcal{A}_o(\mathcal{A}_o\overline{e})\rangle_{L_2,g} + \langle \mathcal{A}_o(\mathcal{A}_o\overline{e}), \overline{e}\rangle_{L_2,g} = 2\langle \overline{e}, \mathcal{A}_o^2\overline{e}\rangle_{L_2,g}.$$

Using the Sobolev embedding theorem [2] (or equivalently the definition of the domain of the operator and integration by parts along with Friedrich's inequality [4]), one may show that

$$2\langle \overline{e}, \mathcal{A}_o^2\overline{e}\rangle_{L_2,g} \geq 2c_1\|\overline{e}\|^2_{L_2,g} \geq 0$$

for some positive c_1 which is related to the embedding constant c in (2.4). The third and fifth terms are

$$\begin{aligned} &-\left\langle \overline{e}, \dot{\xi}^s \frac{\partial(\mathcal{L}(\xi^s)\mathcal{C}(\xi^s))}{\partial \xi^s}\overline{e}\right\rangle_{L_2,g} - \left\langle \dot{\xi}^s \frac{\partial(\mathcal{L}(\xi^s)\mathcal{C}(\xi^s))}{\partial \xi^s}\overline{e}, \overline{e}\right\rangle_{L_2,g} \\ &\quad= -2\left\langle \overline{e}, \dot{\xi}^s \frac{\partial(\mathcal{L}(\xi^s)\mathcal{C}(\xi^s))}{\partial \xi^s}\overline{e}\right\rangle_{L_2,g}, \end{aligned}$$

and, for γ any positive gain, can be made positive by the choice

$$\dot{\xi}_i^s = -\gamma\left\langle \frac{\partial(\mathcal{L}_i(\xi_i^s)\mathcal{C}_i(\xi_i^s))}{\partial \xi_i^s}\overline{e}, \overline{e}\right\rangle_{L_2,g}, \quad i = 1, 2, \ldots, m. \tag{3.13}$$

Finally, we examine the last four terms in the expression for $\dot{V}$. Using similar arguments as made above, we have

(3.14)
$$\begin{aligned} &\langle \mathcal{L}(\xi^s)\left(\mathcal{C}(\xi^s)(\overline{x}-x) - \mathcal{D}v\right), \mathcal{A}_o\overline{e}\rangle_{L_2,g} + \langle \overline{e}, \mathcal{A}_o\mathcal{L}(\xi^s)\left(\mathcal{C}(\xi^s)(\overline{x}-x) - \mathcal{D}v\right)\rangle_{L_2,g} \\ &\quad+ \langle \mathcal{A}_o\mathcal{L}(\xi^s)\left(\mathcal{C}(\xi^s)(\overline{x}-x) - \mathcal{D}v\right), \overline{e}\rangle_{L_2,g} + \langle \mathcal{A}_o\overline{e}, \mathcal{L}(\xi^s)\left(\mathcal{C}(\xi^s)(\overline{x}-x) - \mathcal{D}v\right)\rangle_{L_2,g} \\ &= 2\Big(\langle \mathcal{L}(\xi^s)\left(\mathcal{C}(\xi^s)(\overline{x}-x) - \mathcal{D}v\right), \mathcal{A}_o\overline{e}\rangle_{L_2,g} + \langle \overline{e}, \mathcal{A}_o\mathcal{L}(\xi^s)\left(\mathcal{C}(\xi^s)(\overline{x}-x) - \mathcal{D}v\right)\rangle_{L_2,g}\Big) \\ &= 2\langle (\mathcal{A}_o + \mathcal{A}_o^*)\overline{e}, \mathcal{L}(\xi^s)\left(\mathcal{C}(\xi^s)(\overline{x}-x) - \mathcal{D}v\right)\rangle_{L_2,g} \\ &= 2\Big(\langle (\mathcal{A}_o + \mathcal{A}_o^*)\overline{e}, \mathcal{L}(\xi^s)\mathcal{C}(\xi^s)(\overline{x}-x)\rangle_{L_2,g} - \langle (\mathcal{A}_o + \mathcal{A}_o^*)\overline{e}, \mathcal{L}(\xi^s)\mathcal{D}v\rangle_{L_2,g}\Big). \end{aligned}$$

In the noise-free setting ($v(t) = 0$ and $w(t) = 0$ for all $t \geq 0$), we would have $\overline{x} = x$ and, hence, $\overline{e} = e$. The last four terms are all zero. In this case, we have

$$\begin{aligned} \dot{V} &\leq -2\left\{(1+c_1)\|\overline{e}\|^2_{L_2,g} + \gamma\left(\left\langle \tfrac{\partial(\mathcal{L}(\xi^s)\mathcal{C}(\xi^s))}{\partial \xi^s}\overline{e}, \overline{e}\right\rangle_{L_2,g}\right)^2\right\} \\ &\leq -2(1+c_1)\|\overline{e}\|^2_{L_2,g} \leq -c_2 V, \end{aligned}$$

where $c_2 > 0$ is a constant. The last inequality follows from the application of the $\mathcal{V} \to \mathcal{V}^*$-boundedness property (A1). Hence, we see that the time derivative of V is negative definite. Under the control law (3.13), the error $\overline{e}$ is guaranteed to converge to zero. Within the set of all possible choices of $\mathcal{L}$ (say through a Luenberger or a Kalman filter design) that render the observer dynamics stable, we can further dictate the motion in an attempt to improve the state estimate. The control law (3.13) is essentially a gradient-type control law that seeks to improve the state estimate beyond the capability of a static set of sensors.

Let us now consider the case where we have nonzero process and measurement noise signals. First, consider the dynamics of the error between the idealized process and the actual process. A simple computation gives

$$\frac{d}{dt}(x-\overline{x}) = \mathcal{A}(x-\overline{x}) + \mathcal{B}_1 w. \tag{3.15}$$

Using the fact that the operator $\mathcal{A}$ is the infinitesimal generator of an exponentially stable C_0 semigroup and the L_2-boundedness of $\mathcal{B}_1 w$ (made in Theorem 3.1), well-posedness of (3.15) immediately follows [75]. In fact, we have asymptotic convergence of $x-\overline{x}$ with respect to the $\mathcal{X}$ norm. Given the well-posedness of the above equation, we apply the triangle inequality to obtain $\dot{V} \leq -c_9 V + c_5 + c_7 \|v\|^2$. In the above, all constants are positive and are found by successive application of the $\mathcal{V} \to \mathcal{V}^*$-boundedness of the operator $\mathcal{A}_o$. This shows that V converges to the residual set bounded by

$$\frac{c_5 + c_7 \|v\|^2}{c_9}.$$

Since v is bounded, we readily see that V is bounded. While this does not mean (norm) convergence of the error $\overline{e}$ to zero, in the noisy case, the control law (3.13) drives the error to a neighborhood of zero, which furnishes stability in the sense of Lyapunov.

However, we are interested in the true error e. To show that e converges to a neighborhood of zero, we note that

$$\|\widehat{x}(t)-x(t)\|_{L_2,g} = \|\widehat{x}(t)-\overline{x}(t)+\overline{x}(t)-x(t)\|_{L_2,g} \leq \|\widehat{x}(t)-\overline{x}(t)\|_{L_2,g}+\|\overline{x}(t)-x(t)\|_{L_2,g}.$$

We have already argued that $\overline{x}-x$ converges asymptotically to zero, and that $\overline{e}=\widehat{x}-x$ converges to a neighborhood of zero. Hence, the state estimate $\widehat{x}$ converges to a neighborhood of the true state x. This neighborhood is a function of the bounds on the noise. This gives the following theorem.

THEOREM 3.2. *The control law* (3.13) *with $\mathcal{L}(\xi^s) = \mathcal{C}^*(\xi^s)$ drives the state estimation error e governed by* (3.6) *to a neighborhood of zero as time goes to infinity. In the noise-free case, the state estimation error converges to zero asymptotically.*

REMARK 3.3. *The benefits of the choice $\mathcal{L}(\xi^s) = \mathcal{C}^*(\xi^s)$ are twofold:* (i) *It simplifies the observer gain design by avoiding the solution to either Lyapunov or Riccati operator equations, and* (ii) *it minimizes the computational complexity due to the gradient of both $\mathcal{L}(\xi^s)$ and $\mathcal{C}(\xi^s)$ in* (3.13). *With the above choice only the gradient of $\mathcal{C}(\xi^s)$ with respect to ξ^s is required. In this case, the control law is simply written as*

$$\dot{\xi}_i^s = -2\gamma \left\langle \frac{\partial \mathcal{C}_i(\xi_i^s)}{\partial \xi_i^s}\overline{e}, \mathcal{C}_i(\xi_i^s)\overline{e} \right\rangle_{L_2,g}, \quad i=1,2,\ldots,m. \tag{3.16}$$

4. Numerical results. We simulated the PDE in (2.1) with Dirichlet boundary conditions, having initial conditions $x(0,\xi)=\sin(\frac{\pi\xi}{\ell})e^{-7\xi^2}$ and $\widehat{x}(0,\xi)=0$ as depicted in Figure 4.1. It should be noted that for this initial condition, the "bulk" of the initial state error is in the interval $[0, 0.6\ell]$ and one expects that the sensors be moving in this region as they will be collecting more useful information.

For the specific PDE, the embedding constant is chosen as $c=\pi^{-1}$ [4]. The parameters in the elliptic operator are taken to be $a_1 = 5\times 10^{-3}$, $a_2 = 1.5\times 10^{-1}$,

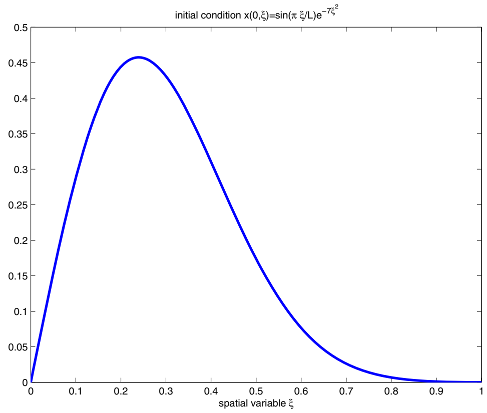


FIG. 4.1. *Distribution of the initial condition $x(0,\xi)$ considered in the simulation studies.*

$a_3 = 3 \times 10^{-3}$, and the length of the spatial domain was taken to be $\ell = 1$. The spatial support of the sensing devices was chosen as $\Delta\xi = \ell/10$.

We approximate (2.1) using linear B-splines [46]. For $n = 1, 2, \ldots$, let $\{\varphi_i^n\}_{i=0}^n$ be the standard B-splines on the interval $[0, \ell]$, defined with respect to the uniform mesh $\{0, \frac{\ell}{n}, \ldots, \ell\}$,

$$\varphi_i^n(\xi) = \begin{cases} 1 - \left|\frac{n\xi}{\ell} - i\right|, & \xi \in \left[\frac{(i-1)\ell}{n}, \frac{(i+1)\ell}{n}\right], \\ 0, & \xi \in [0,\ell] \setminus \left[\frac{(i-1)\ell}{n}, \frac{(i+1)\ell}{n}\right]. \end{cases}$$

We consider a sequence of finite dimensional spaces $\mathcal{X}^n = \operatorname{span}\{\varphi_i^n\}_{i=1}^{n-1}$ and, for each $n = 1, 2 \ldots$, let P_n be the orthogonal projection of $\mathcal{V} = H_0^1(0,\ell)$ into $\mathcal{X}^n$. We let $X^n(t) \in \mathbb{R}^{n-1}$ be the coordinate vector for $x^n(t)$ with respect to the basis $\{\varphi_i^n\}_{i=1}^{n-1}$,

$$x^n(t) = P_n x(t) = \sum_{i=1}^{n-1} X_i^n(t)\varphi_i^n(\xi).$$

Let $\widehat{x}^n(t) \in \mathbb{R}^{n-1}$ be the coordinate vector for the finite dimensional approximation of $\widehat{x}(t)$, with $\widehat{x}^n(t) = \sum_{j=1}^{n-1} \widehat{X}_j(t)\varphi_j^n(\xi)$. We denote by M^n the Gram matrix corresponding to $\{\varphi_i^n\}_{i=1}^{n-1}$, and thus we obtain

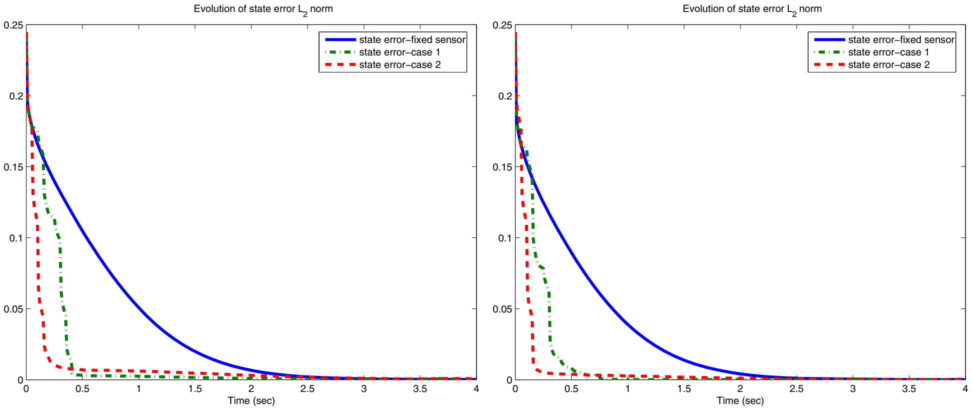
$$M^n = [M_{ij}^n] = \left[\int_0^\ell \varphi_i^n(\xi)\varphi_j^n(\xi)\, d\xi\right].$$

Additionally, we let K^n, L^n be the $(n-1) \times (n-1)$ matrices

$$K^n = [K_{ij}^n] = \left[\int_0^\ell d\varphi_i^n(\xi) d\varphi_j^n(\xi)\, d\xi\right], \qquad L^n = [L_{ij}^n] = \left[\int_0^\ell d\varphi_i^n(\xi)\varphi_j^n(\xi)\, d\xi\right].$$

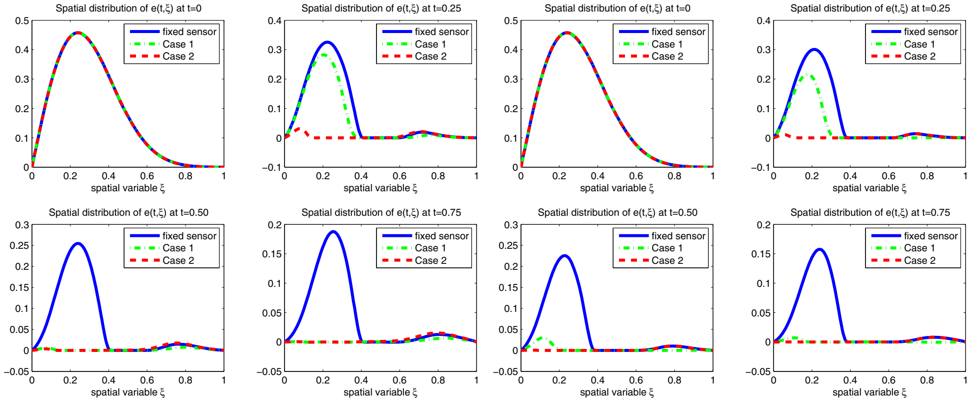
The matrix representations of (2.1) and (3.1) then become

$$\begin{aligned} M^n \dot{X}^n(t) &= \left(-a_1 K^n - a_2 L^n - a_3 M^n\right) X^n(t) + B_1^n(t) w(t) + B_2^n u(t), \\ M^n \dot{\widehat{X}}^n(t) &= \left(-a_1 K^n - a_2 L^n - a_3 M^n\right) \widehat{X}^n(t) \\ &\quad + B_2^n u(t) + L^n(\xi^s)\Big(y^n(t;\xi^s) - C^n(\xi^s)\widehat{X}^n(t)\Big). \end{aligned}$$



(a) using one sensor (b) using two sensors

FIG. 4.2. *Evolution of state errors norms.*

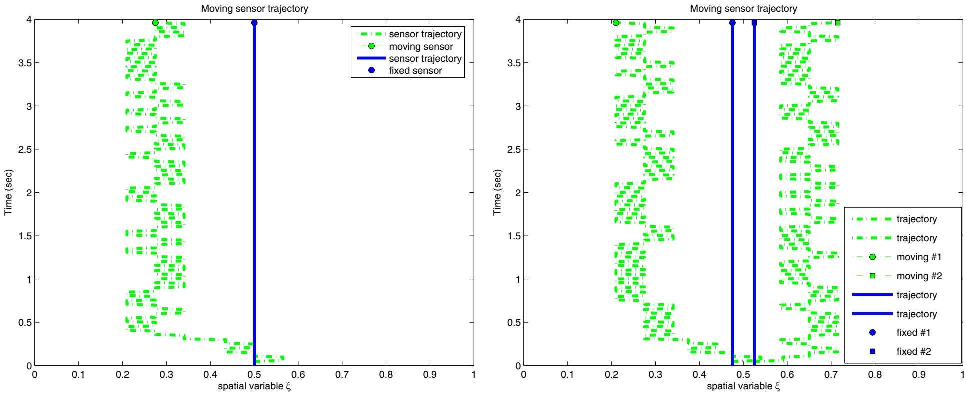


(a) using one sensor (b) using two sensors

FIG. 4.3. *Spatial distributions of the error $e(t, \xi)$ at different time instances.*

In both cases, the filter gain $\mathcal{L}(\xi^s)$ was taken to be equal to $\mathcal{C}^*(\xi^s)$. We simulated the N-dimensional system with 80 basis elements [46] that preserve exponential detectability [6]. The computations were carried out via codes written in MATLAB run on a dual processor DELL workstation (Xeon 2.8GHz, 2 × 2GB). The resulting finite dimensional system of ordinary differential equations (ODEs) was integrated using the stiff ODE solver from the MATLAB ODE library, routine ode23s based on a 4th Runge–Kutta scheme. All spatial integrals were computed numerically via a composite two point Gauss–Legendre quadrature rule [4].

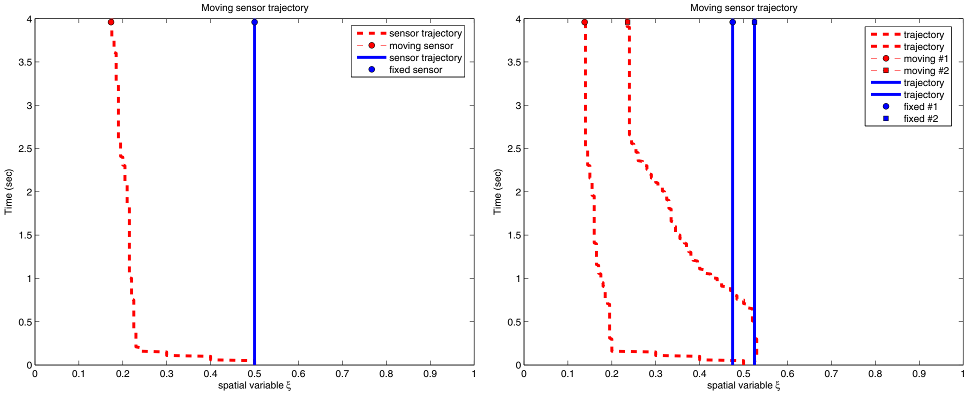
The evolution of the state error norm for the mobile and fixed-sensor cases is presented in Figure 4.2. It is observed that when the sensor is allowed to move within the spatial domain, the estimation error converges to zero faster. This is true for both proposed guidance policies: *localized measurement error* and *global estimation error*. The spatial distribution of the state error at different time instances is depicted in Figure 4.3, where one can once again observe the ability of a mobile sensor to estimate the state faster. In both the case of one and two mobile sensors, Case 2 (based on the



(a) using one sensor

(b) using two sensors

FIG. 4.4. *Case* 1. *Sensor trajectory: moving (dashed lines), fixed (solid lines).*



(a) using one sensor

(b) using two sensors

FIG. 4.5. *Case* 2. *Sensor trajectory: moving (dotted lines), fixed (solid lines).*

global estimation error) tends to give better results than Case 1 (based on the localized measurement error). The sensor trajectories for both cases are presented in Figures 4.4 and 4.5. The initial position for the single sensor case was chosen as $\xi_1^s(0) = 0.5\ell$, and initial conditions for the two-sensor case were chosen as $\xi_1^s(0) = 0.475\ell$ and $\xi_1^s(0) = 0.525\ell$. By examination of the initial condition, and hence the initial condition of the estimation error, it is observed that the guidance policies send the sensor(s) to the region of largest spatial error.

These results clearly validate the basic premise of this paper. Namely, a set of mobile sensors moving according to either one of two guidance policies proposed in this paper will perform better than a set of static sensors located at the initial locations of the mobile sensors.

5. Summary and concluding remarks. In this paper we have considered the problem of controlling a network of fully connected, sensor-equipped vehicles to estimate a spatially distributed process described by a linear PDE. The process was assumed to be driven by a zero mean Gaussian noise and the goal was to improve state estimation via the use of spatially distributed mobile sensors. By utilizing the resulting state estimation error, two guidance policies for the mobile sensors were proposed. The first guidance policy seeks to have each agent minimize the infinity norm of the state estimation error over the sensory domain of the associated sensor. Implicitly imbedded into the sensor guidance policy was a velocity requirement in the sense of moving a given sensor to the spatial location within the domain that had the largest deviation of the estimation error. Such a guidance policy rendered the error system a hybrid one having system operators that generate an exponentially stable C_0 semigroup and forcing terms that satisfied an L_2 bound.

The second guidance policy seeks to have each agent minimize the L_2 norm of the global estimation error over the entire domain $\mathcal{D}$. A Lyapunov-based argument was used to show that the L_2 state estimation error associated with the nominal process monotonically decreases until the error is zero within the ranges of all sensors in the network. Simulation studies implementing and comparing the two proposed control policies were provided. The simulations show that moving the sensors according to the proposed control laws is advantageous to not moving the sensors at all. While both the static and mobile cases eventually achieve zero estimation error, the mobile sensors converge faster than the static network.

Both methods required that the filter gain be equal to the adjoint of the measurement operator. Such a choice significantly simplifies the observer gain design and, more importantly, minimizes the computational requirements required when one solves an optimal filter problem via the solution to associated filter differential Riccati equations.

While the current goal was to estimate the process state efficiently, the more interesting case of utilizing mobile sensors would be to detect spatiotemporally varying disturbances and moving sources that may represent contamination or intrusion. Preliminary work on such a case that utilizes the above methods within the abstract theory of infinite dimensional systems has recently been considered in [26] for simple detection of a moving source within a two-dimensional spatial domain and in [27, 28] for the integrated state estimation, intrusion detection, and containment. While the proposed framework easily allows for two- and three-dimensional diffusion-advection processes governed by elliptic operators, a major challenge results in the numerical implementation as the dimension of the finite dimensional representation of the process increases polynomially. However, in this case efficient model-reduction schemes that are based on Karhunen–Loève expansions may be incorporated in order to allow for real-time feasibility. Such a task is currently being undertaken by the first author.

In the current work, vehicle cooperation is in the sense that spatial information is shared between full-connected (communicationwise) vehicles. This sharing of information allows for the coordination of the motion in order to achieve satisfactory estimates of the field over a given domain. There are two open questions that remain to be addressed regarding increased vehicle cooperation. The first involves relaxing the communication full-connectedness assumption. Methods recently developed by the second author (see [89]) that guarantee satisfactory domain coverage under arbitrary intermittent communication structures, with decentralized decision making, can be applied to the two strategies developed in this paper, especially the second strategy. Other communication considerations, such as time delays, fading channels, partially connected and dynamic communication structures, will also be the focus of future work by the authors, as well as distributed processes governed by nonlinear dynamics. However for the abstract framework considered here, collision avoidance takes additional importance as one must restrict the motion of the mobile sensors within the set of admissible sensor locations in order to guarantee observability.

REFERENCES

[1] E. U. Acar and H. Choset, *Sensor-based coverage of unknown environments: Incremental construction of Morse decompositions*, Internat. J. Robotics Res., 21 (2002), pp. 345–366.

[2] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[3] M. I. Asensio and L. Ferragut, *On a wildland fire model with radiation*, Internat. J. Numer. Methods Engrg, 54 (2002), pp. 137–157.

[4] O. Axelsson and V. A. Barker, *Finite Element Solutions of Boundary Value Problems*, Academic Press, Orlando, FL, 1984.

[5] J. Baillieul and P. J. Antsaklis, *Control and communication challenges in networked real-time systems*, Proceedings IEEE, 95 (2007), pp. 9–28.

[6] H. T. Banks, R. C. Smith, and Y. Wang, *Smart Material Structures: Modeling, Estimation and Control*, Wiley-Masson, New York, 1996.

[7] J. S. Baras and A. Bensoussan, *Optimal sensor scheduling in nonlinear filtering of diffusion processes*, SIAM J. Control Optim., 27 (1989), pp. 786–813.

[8] V. Barbu, *Nonlinear Semigroups and Differential Equations in Banach Space*, Noordhoff, Leyden, The Netherlands, 1976.

[9] A. F. Bennett, *Inverse Modeling of the Ocean and Atmosphere*, Cambridge University Press, Cambridge, UK, 2002.

[10] A. G. Butkovskiĭ, *Theory of mobile control*, Automat. Remote Control, 1979, no. 6, pp. 29–41 (in Russian).

[11] A. G. Butkovskiĭ and E. I. Pustyl'nikova, *Theory of mobile control of distributed parameter systems*, Automat. Remote Control, 1980, no. 6, pp. 5–13 (in Russian).

[12] A. G. Butkovskij and L. M. Pustyl'nikov, *Mobile Control of Distributed Parameter Systems*, Ellis Horwood Limited, Chichester, UK, 1987.

[13] J. Caffrey, R. Govindan, E. Johnson, B. Krishnamachari, S. Masri, G. Sukhatme, K. Chintalapudi, K. Dantu, S. Ranwala, A. Sritharan, N. Xu, and M. Zuniga, *Networked sensing for structural health monitoring*, in Proceedings of the 4th International Workshop on Structural Control, Columbia University, New York, 2004, pp. 57–66.

[14] C. G. Cassandras and W. Li, *Sensor networks and cooperative control*, European J. Control, 11 (2005), pp. 436–463.

[15] S. Chakravorty, *Design and Optimal Control of Multi-Spacecraft Interferometric Imaging Systems*, Ph.D. thesis, Aerospace Engineering, University of Michigan, Ann Arbor, MI, 2004.

[16] H. Choset, *Coverage for robotics: A survey of recent results*, Ann. Math. Artificial Intell., 31 (2001), pp. 113–126.

[17] T. H. Chung, V. Gupta, J. W. Burdick, and R. M. Murray, *On a decentralized active sensing strategy using mobile sensor platforms in a network*, in Proceedings of the IEEE Conference on Decision and Control, Paradise Island, Bahamas, 2004, pp. 1914–1919.

[18] J. Cortés, S. Martínez, and F. Bullo, *Spatially-distributed coverage optimization and control with limited-range interactions*, ESAIM Control Optim. Calc. Var., 11 (2005), pp. 691–719.

[19] J. Cortés, S. Martínez, T. Karatus, and F. Bullo, *Coverage control for mobile sensing networks*, IEEE Trans. Robotics Automat., 20 (2004), pp. 243–255.

[20] R. F. Curtain and A. Ichikawa, *Optimal location of sensors for filtering for distributed systems*, in Distributed Parameter Systems: Modelling and Identification (Proc. IFIP Working Conf., Rome, 1976), Lecture Notes in Control and Inform. Sci., Springer-Verlag, Berlin, 1978, pp. 236–255.

[21] R. F. Curtain and H. J. Zwart, *An Introduction to Infinite Dimensional Linear Systems Theory*, Texts in Appl. Math. 21, Springer-Verlag, Berlin, 1995.

[22] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology, Vol.* 2: *Functional and Variational Methods*, Springer-Verlag, Berlin, Heidelberg, New York, 2000.

[23] M. A. Demetriou, *Activation policy of smart controllers for flexible structures with multiple actuator/sensor pairs*, in Proceedings of the Fourteenth International Symposium on Mathematical Theory of Networks and Systems (MTNS 2000), Perpignan, France, 2000, CD-ROM.

[24] M. A. Demetriou, *Vibration control of flexible structures using an optimally moving actuator*, in Proceedings of the 14th ASCE Engineering Mechanics Division Conference, University of Texas at Austin, Austin, TX, 2000, CD-ROM.

[25] M. A. Demetriou, *Integrated actuator/sensor placement and hybrid controller design of flexible structures under worst case spatiotemporal disturbance variations*, J. Intell. Material Systems Structures, 15 (2005), pp. 901–931.

[26] M. A. Demetriou, *Power management of sensor networks for detection of a moving source in 2-D spatial domains*, in Proceedings of the American Control Conference, Minneapolis, MN, 2006, pp. 1144–1149.

[27] M. A. Demetriou, *Detection and containment policy of moving source in 2-d diffusion processes using sensor/actuator network*, in Proceedings of the European Control Conference, Kos, Greece, 2007.

[28] M. A. Demetriou, *Process estimation and moving source detection in 2-D diffusion processes by scheduling of sensor networks*, in Proceedings of the American Control Conference, New York, 2007, pp. 3432–3437.

[29] M. A. Demetriou and J. Borggaard, *Optimization of a joint sensor placement and robust estimation scheme for distributed parameter processes subject to worst spatial disturbance distributions*, in Proceedings of the American Control Conference, Boston, MA, 2004, pp. 2239–2244.

[30] M. A. Demetriou and O. V. Iftime, *Finite horizon optimal control of switched distributed parameter systems with moving actuators*, in Proceedings of the American Control Conference, Portland, OR, 2005, pp. 3912–3917.

[31] M. A. Demetriou and N. Kazantzis, *Performance enhancement of controlled diffusion processes by moving actuators*, in Proceedings of the Fifteenth International Symposium on Mathematical Theory of Networks and Systems, University of Notre Dame, Notre Dame, IN, 2002, CD-ROM.

[32] M. A. Demetriou and N. Kazantzis, *Compensation of spatiotemporally varying disturbances in nonlinear transport processes via actuator scheduling*, Internat. J. Robust Nonlinear Control, 14 (2004), pp. 191–197.

[33] M. A. Demetriou and N. Kazantzis, *A new actuator activation policy for performance enhancement of controlled diffusion processes*, Automatica, 40 (2004), pp. 415–421.

[34] M. A. Demetriou and N. Kazantzis, *A new integrated output feedback controller synthesis and collocated actuator/sensor scheduling framework for distributed parameter processes*, Comput. Chem. Engrg., 29 (2005), pp. 867–876.

[35] M. A. Demetriou, A. Paskaleva, O. Vayena, and H. Doumanidis, *Scanning actuator guidance scheme in a 1-d thermal manufacturing process*, IEEE Trans. Control Systems Technology, 11 (2003), pp. 757–764.

[36] D. V. Dimarogonas, S. G. Loizou, K. J. Kyriakopoulos, and M. M. Zavlanos, *A feedback stabilization and collision avoidance scheme for multiple independent non-point agents*, Automatica, 42 (2006), pp. 229–243.

[37] Z. Drezner, *Facility Location: A Survey of Applications and Methods*, Springer, New York, 1995.

[38] Q. Du, V. Faber, and M. Gunzburger, *Centroidal Voronoi tessellations: Applications and algorithms*, SIAM Rev., 41 (1999), pp. 637–676.

[39] J.-L. Dupuy and M. Larini, *Fire spread through a porous forest fuel bed: A radiative and convective model including fire-induced flow effects*, Internat. J. Wildland Fire, 9 (1999), pp. 155–172.

[40] A. Ganguli, J. Cortés, and F. Bullo, *Maximizing visibility in nonconvex polygons: Nonsmooth analysis and gradient algorithm design*, SIAM J. Control Optim., 45 (2006), pp. 1657–1679.

[41] A. Ganguli, S. Susca, S. Martínez, F. Bullo, and J. Cortés, *On collective motion in sensor networks: Sample problems and distributed algorithms*, in Proceedings of the 44th IEEE Conference on Decision and Control, Seville, Spain, 2005, pp. 4239–4244.

[42] V. Giordano, P. Ballal, F. Lewis, B. Turchiano, and J. B. Zhang, *Supervisory control of mobile sensor networks: Math formulation, simulation, and implementation*, IEEE Trans. Systems Man Cybernet. B Cybernet., 36 (2006), pp. 806–819.

[43] B. Grocholsky, *Information-Theoretic Control of Multiple Sensor Platforms*, Ph.D. thesis, The University of Sydney, Sydney, Australia, 2002.

[44] V. Gupta, T. H. Chung, B. Hassibi, and R. M. Murray, *On a stochastic sensor selection algorithm with applications in sensor scheduling and sensor coverage*, Automatica, 42 (2006), pp. 251–260.

[45] K. H. Hoffmann, P. Knabner, and W. Seifert, *Adaptive methods for parameter identification in ground water hydrology*, Adv. Water Res., 14 (1991), pp. 220–239.

[46] K. Höllig, *Finite Element Methods with B-Splines*, Frontiers in Appl. Math. 26, SIAM, Philadelphia, 2003.

[47] I. I. Hussein, *Motion Planning for Multi-Spacecraft Interferometric Imaging Systems*, Ph.D. thesis, University of Michigan, Ann Arbor, MI, 2005.

[48] I. I. Hussein, *A Kalman filter-based control strategy for dynamic coverage control*, in Proceedings of the American Control Conference, New York, 2007, pp. 3271–3276.

[49] I. I. Hussein and M. A. Demetriou, *Estimation of distributed processes using mobile spatially distributed sensors*, in Proceedings of the American Control Conference, New York, 2007, pp. 2756–2761.

[50] I. I. Hussein and D. Stipanović, *Effective coverage control for mobile sensor networks*, in Proceedings of the IEEE Conference on Decision and Control, San Diego, 2006, pp. 2747–2752.

[51] I. I. Hussein and D. Stipanović, *Effective coverage control for mobile sensor networks with guaranteed collision avoidance*, IEEE Trans. Control Systems Technology, 15 (2007), pp. 642–657.

[52] I. I. Hussein and D. Stipanović, *Effective coverage control using dynamic sensor networks with flocking and guaranteed collision avoidance*, in Proceedings of the American Control Conference, New York, 2007, pp. 3420–3425.

[53] O. V. Iftime and M. A. Demetriou, *Optimal control for switched distributed parameter systems with application to the guidance of a moving actuator*, in Proceedings of the 16th IFAC World Congress, Prague, 2005, CD-ROM.

[54] E. Kalnay, *Atmospheric Modeling, Data Assimilation and Predictability*, Cambridge University Press, Cambridge, UK, 2003.

[55] A. Khapalov, *L^∞-exact observability of the heat equation with scanning pointwise sensor*, SIAM J. Control Optim., 32 (1994), pp. 1037–1051.

[56] A. Khapalov, *Mobile point controls versus locally distributed ones for the controllability of the semilinear parabolic equation*, SIAM J. Control Optim., 40 (2001), pp. 231–252.

[57] G. Leitmann, *Guaranteed avoidance strategies*, J. Optim. Theory Appl., 32 (1980), pp. 569–576.

[58] G. Leitmann and J. Skowronski, *Avoidance control*, J. Optim. Theory Appl., 23 (1977), pp. 581–591.

[59] F. Lewis, *Wireless sensor networks*, in Smart Environments: Technologies, Protocols, and Applications, John Wiley, New York, 2004, pp.

[60] W. Li and C. G. Cassandras, *Distributed cooperative coverage control of sensor networks*, in Proceedings of the IEEE Conference on Decision and Control, Seville, Spain, pp. 2542–2547.

[61] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.

[62] S. Lloyd, *Least squares quantization in pcm*, IEEE Trans. Inform. Theory, 28 (1982), pp. 129–137.

[63] J. Mandel, M. Chen, L. P. Franca, C. Johns, A. Puhalskii, J. L. Coen, C. C. Douglas, R. Kremens, A. Vodacek, and W. Zhao, *A Note on dynamic data driven wildfire modeling*, in Proceedings of ICCS 2004, Lecture Notes in Comput. Sci. 3038, Springer, Berlin, 2004, pp. 725–731.

[64] S. Martínez and F. Bullo, *Optimal sensor placement and motion coordination for target tracking*, Automatica, 42 (2006), pp. 661–668.

[65] S. Martínez, F. Bullo, J. Cortés, and E. Frazzoli, *On synchronous robotic networks—part* ii: *Time complexity of rendezvous and deployment algorithms*, IEEE Trans. Automat. Control, 52 (2007), pp. 2214–2226.

[66] S. Martínez, F. Bullo, J. Cortés, and E. Frazzoli, *On synchronous robotic networks—part* i: *Models, tasks and complexity*, IEEE Trans. Automat. Control, 52 (2007), pp. 2199–2213.

[67] S. Martínez, J. Cortés, and F. Bullo, *Motion coordination with distributed information*, IEEE Control Systems Mag., 27 (2007), pp. 75–88.

[68] R. Mehra, *Optimization of measurement schedules and sensor designs for linear dynamic systems*, IEEE Trans. on Automatic Control, 21 (1976), pp. 55–64.

[69] Y. Mostofi, T. Chung, R. Murray, and J. Burdick, *Communication and sensing trade offs in decentralized mobile sensor networks: A cross-layer design approach*, in Proceedings of the 4th International Conference on Information Processing in Sensor Networks, Los Angeles, CA, 2005, pp. 118–125.

[70] Y. Mostofi and R. Murray, *Distributed sensing and estimation under communication constraints*, in Proceedings of the IEEE Conference on Decision and Control, San Diego, CA, 2006, pp. 1013–1018.

[71] *NASA Origins Program*, http://origins.jpl.nasa.gov/index1.html (2008).

[72] A. Nehorai, B. Porat, and E. Paldi, *Detection and localization of vapor-emitting sources*, IEEE Trans. Signal Process., 43 (1995), pp. 243–253.

[73] A. Okabe and A. Suzuki, *Locational optimization problems solved through Voronoi diagrams*, European J. Oper. Res., 98 (1997), pp. 445–456.

[74] S. Omatu and J. H. Seinfeld, *Distributed Parameter Systems: Theory and Applications*, Oxford University Press, NY, 1989.

[75] J. Ooostveen, *Strongly Stabilizable Distributed Parameter Systems*, Frontiers in Appl. Math. 20, SIAM, Philadelphia, 2000.

[76] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[77] J. C. Robinson, *Infinite-Dimensional Dynamical Systems: An Introduction to Dissipative Parabolic PDEs and the Theory of Global Attractors*, Cambridge University Press, Cambridge, UK, 2001.

[78] J. H. Seinfeld, *Atmospheric Chemistry and Physics of Air Pollution*, John Wiley & Sons, New York, 1995.

[79] J. H. Seinfeld and S. N. Pandis, *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*, Wiley-Interscience, New York, 1997.

[80] R. E. Showalter, *Hilbert Space Methods for Partial Differential Equations*, Pitman, London, 1977.

[81] Z. Song, Y. Chen, J. Liang, and D. Uciński, *Optimal mobile sensor motion planning under nonholonomic constraints for parameter estimation of distributed systems*, in Proceedings of the 2005 IEEE IRS/RSJ International Conference on Intelligent Robots and Systems, Edmonton, Alberta, Canada, 2005, pp. 1505–1510.

[82] D. M. Stipanović, P. F. Hokayem, M. W. Spong, and D. D. Šiljak, *Cooperative avoidance control for multi-agent systems*, ASME J. Dynam. Systems Measurement Control, 129 (2007), pp. 699–707.

[83] D. M. Stipanović, S. Shankaran, and C. J. Tomlin, *Multi-agent avoidance control using an M-matrix property*, Electron. J. Linear Algebra, 12 (2005), pp. 64–72.

[84] S. Susca, S. Martínez, and F. Bullo, *Monitoring environmental boundaries with a robotic sensor network*, IEEE Trans. Control Systems Technology, (2008), pp. 288–296.

[85] H. Tanabe, *Equations of Evolution*, Pitman, London, 1979.

[86] D. Uciński, *Sensor network design for parameter estimation of distributed systems using nonsmooth optimality criteria*, in Proceedings of the IEEE Conference on Decision and Control, Seville, Spain, 2005, pp. 3152–3157.

[87] D. Uciński and Y. Chen, *Time-optimal path planning of moving sensors for parameter estimation of distributed systems*, in Proceedings of the IEEE Conference on Decision and Control, Seville, Spain, 2005, pp. 5257–5262.

[88] D. Uciński and J. Korbicz, *Path planning for moving sensors in parameter estimation of distributed systems*, in Proceedings of the 1st Workshop on Robot Motion and Control, Kiekrz, Poland, 1999, pp. 273–278.

[89] Y. Wang and I. Hussein, *Awareness coverage control over large scale domains with intermittent communications*, in Proceedings of the American Control Conference, Seattle, WA, 2008, pp. 4370–4375.

[90] J. Wloka, *Partial Differential Equations*, Cambridge University Press, Cambridge, UK, 1987.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 1, pp. 292–320

© 2009 Society for Industrial and Applied Mathematics

A GEOMETRIC OPTIMIZATION APPROACH TO DETECTING AND INTERCEPTING DYNAMIC TARGETS USING A MOBILE SENSOR NETWORK[*]

SILVIA FERRARI[†], RAFAEL FIERRO[‡], BRENT PERTEET[§], CHENGHUI CAI[†], AND KELLI BAUMGARTNER[†]

Abstract. A methodology is developed to deploy a mobile sensor network for the purpose of detecting and capturing mobile targets in the plane. The sensing-pursuit problem considered in this paper is analogous to the Marco Polo game, in which a pursuer Marco must capture multiple mobile targets that are sensed intermittently, and with very limited information. The competing objectives exhibited by this problem arise in a number of surveillance and monitoring applications. In this paper, the mobile sensor network consists of a set of robotic sensors that must track and capture mobile targets based on the information obtained through cooperative detections. When these detections form a satisfactory target track, a mobile sensor is switched to pursuit mode and deployed to capture the target in minimum time. Since the sensors are installed on robotic platforms and have limited range, the geometry of the platforms and of the sensors' fields-of-view play a key role in obstacle avoidance and target detection. A new cell-decomposition approach is presented to determine the probability of detection and the cost of operating the sensors from the geometric properties of the network and its workspace. The correctness and complexity of the algorithm are analyzed, proving that the termination time is a function of the network parameters and of the number of required detections.

Key words. mobile sensor networks, pursuit-evasion games, coverage, tracking, detection

AMS subject classifications. 49N75, 46N10, 74P20, 93E10

DOI. 10.1137/07067934X

1. Introduction.

The proliferation of reliable low-cost sensors and autonomous vehicles is producing advanced surveillance systems comprised of robotic sensors with a high degree of functionality and reconfigurability. These mobile sensor networks can play a critical role in several application domains, such as landmine detection and identification [37, 14]; monitoring of endangered species [25]; monitoring of urban environments, manufacturing plants, and civil infrastructure; high-confidence medical devices; and intruder and target detection systems. These networks are expected to operate cooperatively and reliably in cluttered dynamic environments with little human intervention. Coordinating such large heterogeneous sensor networks is challenging and requires the development of novel methods of communication, motion control and planning, computation, proactive estimation and sensing, and power management.

[*] Received by the editors January 6, 2007; accepted for publication (in revised form) December 1, 2008; published electronically February 11, 2009.
http://www.siam.org/journals/sicon/48-1/67934.html

[†] Laboratory for Intelligent Systems and Controls, Department of Mechanical Engineering & Materials Science, Duke University, Durham, NC 27708-0300 (sferrari@duke.edu, cc88@duke.edu, kacb@alumni.duke.edu). The work of the first author was supported in part by the Office of Naval Research (Code 321), and by NSF grant ECS CAREER #0448906.

[‡] MARHES Lab, Department of Electrical & Computer Engineering, University of New Mexico, Albuquerque, NM 87131-0001 (rfierro@ece.unm.edu). The work of this author was supported by NSF grants ECCS CAREER #0811347, IIS #0812338, CNS #0709329, by DOE NNSA grant DE-FG52-04NA25590 (through the UNM Manufacturing Engineering Program), and by the U.S. Army Research Office under grant DAAD19-03-1-0142 (through the University of Oklahoma).

[§] Department of Electrical & Computer Engineering, Oklahoma State University, Stillwater, OK 74078-5032 (brent.perteet@gmail.com).

One paradigm common to many sensing applications consists of one or more sensors installed on robotic platforms that must move through an environment to obtain measurements from multiple targets. Most of the research relating sensor measurements to robot motion planning has focused on the effects that the uncertainty in the geometric models of the environment has on the motion strategies of the robot [27, 42, 41, 35, 33]. Hence, considerable progress has been made toward integrating sensor measurements in topological maps [46], and on planning strategies based on only partial or nondeterministic knowledge of the workspace [30, 32]. Coordination of robotic networks and sensor planning approaches have received considerable attention in recent years [51, 11, 3]. One line of research has investigated the extension of motion planning techniques to the problem of sensor placement for achieving coverage of unstructured environments [1, 9] or of a desired visibility space [30, 22]. Obstacle-avoidance motion planners have been effectively modified in [40, 8] to plan the path of mobile sensors for the detection and classification of stationary targets in an obstacle-populated environment. Probabilistic pursuit-evasion strategies to detect and capture intelligent evaders in obstacle-populated environments are described in [48]. In [24] the authors show that a pursuer can detect an arbitrarily fast evader in a polygonal environment using a randomized strategy. It is shown that one evader is guaranteed to be captured by two pursuers in finite time, by solving a *lion and man* problem and assuming that at least one pursuer is as fast as the evader [24].

In this paper, we develop a cell-decomposition methodology to optimize the probability of detection of a mobile sensor network, based on the geometry of the workspace and of the robotic sensors. Cell-decomposition algorithms have previously been employed to represent the obstacle-free configuration of a robot for the purpose of obstacle avoidance [29]. We present a framework for obtaining a decomposition in which observation cells are used to represent sensor configurations that intersect the targets while avoiding polygonal obstacles. The simple philosophy behind this approach is that while the geometry of the robot must not intersect that of an obstacle to avoid collision, the geometry of the sensor's field-of-view (or visibility region) must intersect that of a target to enable a detection. Then, the tracking information is used to determine the probability of detection in the observation cells.

At any given time, the pursuers must also detect new targets, for which there is no available track information. The monitoring of a workspace by means of multiple sensors is typically referred to as coverage. Coverage control for mobile sensors has been treated in [11] using Voronoi diagrams to achieve uniform sensing performance over an area-of-interest. Another well-known coverage problem is the art-gallery or line-of-sight visibility problem, in which multiple sensors are placed such that the targets are in the line-of-sight of at least one sensor in the network [36, 45, 47]. In this paper, we consider a track-coverage formulation in which multiple sensors are deployed to cooperatively detect moving targets traversing the area-of-interest [17]. By this formulation, the probability of detection of undetected targets is obtained for every cell in the decomposition. Then, the control policies that optimize a trade-off of multiple sensing objectives are obtained by searching the robot configuration graph and by performing inner-loop trajectory generation and tracking. The path obtained from the configuration graph is one that maximizes the overall probability of detection and minimizes the distance traveled by the pursuer to detect or capture the targets.

The remainder of the paper is organized as follows. The sensing-pursuit problem is formulated in section 2. The geometric approach used to control the network of robotic sensors to detect and pursue moving targets is presented in section 3. The correctness and performance analysis of the algorithm is presented in section 4. The simulation results obtained are described in section 5.

2. Problem statement and assumptions. We consider a pursuit-evasion game in which N pursuers comprised of robotic sensors attempt to detect and pursue M moving targets. The game takes place in a square area-of-interest $\mathcal{S} \subset \mathbb{R}^2$, with boundary $\partial\mathcal{S}$ and dimensions $L \times L$. $\mathcal{S}$ is populated by n fixed and convex obstacles $\{\mathcal{O}_1, \ldots, \mathcal{O}_n\} \subset \mathcal{S}$. The geometry of the ith pursuer is assumed to be a convex polygon denoted by $\mathcal{A}_i$, with a configuration q_i that specifies its position and orientation with respect to a fixed Cartesian frame $\mathcal{F}_\mathcal{S}$.

The dynamics of the pursuers can be approximated using the nonholonomic unicycle model,

$$\begin{aligned} \dot{x}_p^i &= v_p^i \cos\theta_p^i, \\ \dot{y}_p^i &= v_p^i \sin\theta_p^i, \\ \dot{\theta}_p^i &= \omega_p^i, \end{aligned} \tag{2.1}$$

where $q_i = (x_p^i,\ y_p^i,\ \theta_p^i) \in SE(2)$ and $p_i = [x_p^i \quad y_p^i]^T \in \mathbb{R}^2$ is the position vector of pursuer i (referred to as its centroid). The input to pursuer i is $u_p^i = [v_p^i \quad \omega_p^i]^T$, and $u_p \in \mathcal{U} \subset \mathbb{R}^2$. The set of all pursuers is denoted by $\mathcal{P}$, and $I_\mathcal{P}$ is the index set of $\mathcal{P}$. The set of all targets in $\mathcal{S}$ is denoted by $\mathcal{T}$, where $I_\mathcal{T}$ is the index set of $\mathcal{T}$. The model of the targets is given by

$$\begin{aligned} \dot{x}_\tau^j &= c_{x_\tau}^j, \\ \dot{y}_\tau^j &= c_{y_\tau}^j, \end{aligned} \tag{2.2}$$

where $\tau_j = [x_\tau^j \quad y_\tau^j]^T \in \mathbb{R}^2$ is the position vector of target j and $c_{x_\tau}^j$ and $c_{y_\tau}^j$ are constants. In other words, targets are assumed to move along straight lines

$$y_\tau^j(t) = \frac{c_{y_\tau}^j}{c_{x_\tau}^j} x_\tau^j(t) + y_\tau^j(0) - \frac{c_{y_\tau}^j}{c_{x_\tau}^j} x_\tau^j(0), \tag{2.3}$$

where $\tau_j(0) = (x_\tau^j(0),\ y_\tau^j(0)) \in \partial\mathcal{S}$, and they remain in $\mathcal{S}$ at all $t > 0$. Exceptions to this rule are maneuvers used to avoid an obstacle or another target. The heading angle of a target j is denoted by θ_τ^j; thus $\theta_\tau^j := \arctan(c_{y_\tau}^j, c_{x_\tau}^j)$. The maximum translational speed $V_{p,\tau_{\max}}$ of all sensors and targets is known, and $V_{p_{\max}} > V_{\tau_{\max}}$ [24]. While sensors can move with any speed in $[0, V_{p_{\max}}]$, it is assumed that the speed of every target is uniformly distributed in $[V_{\tau_{\min}}, V_{\tau_{\max}}]$, with $V_{\tau_{\min}} > 0$.

In the sensing problem, the paths of the targets are represented by rays or half-lines, denoted by $\mathcal{R}_\theta^j$, that are unknown a priori. The sensors installed on the robotic platforms are assumed to be isotropic or omnidirectional, and therefore their field-of-view is represented by a disk $\mathcal{D}_i = \mathcal{D}(p_i, r_i) \in \mathcal{S}$ with radius r_i and centered at p_i. The sensor i installed on the robot $\mathcal{A}_i$ has the ability to *detect* the jth target when $\mathcal{D}_i \cap \mathcal{R}_\theta^j \neq \varnothing$. The measurements obtained from each detection can be associated with a particular target using a data-association algorithm (such as [39, 12, 21]), but they may be subject to errors and false alarms. At any time t, the set of detections associated with a target j is denoted by Z_j^t, which symbolizes all measurements of the target positions τ_j obtained since the onset of the game t_0; e.g., $Z_j^t = \{z_j(t_1), z_j(t_2), \ldots, z_j(t_l)\}$. Since the sensors produce few individual observations for each moving target (e.g., due to their limited range) and are subject to frequent false alarms, the approach known as *track-before-detect* [50] is used, in which a set of k spatially distributed sensor detections are used to estimate the *target track*, $\mathcal{R}_\theta^j$, from Z_j^t, before declaring a positive detection. Every track may be updated every time a new measurement becomes available from the target. Once a target track has been formed from at least k sensor detections that are obtained at different moments in time, an upper-level controller declares the target positively detected and deploys a pursuer to capture it. The inputs to the pursuers take into account the information available from all targets, $Z^t = \{Z_j^t \mid j \in I_\mathcal{T}\}$, in order to optimize their sensing and pursuit performance.

Let e_{ji} be the Euclidean distance from the jth target position, τ_j, to the closest pursuer; i.e., $e_{ji} = \min d(\tau_j, p_i)\ \forall i \in I_\mathcal{P}$. Then the pursuer i is said to *capture* the target j when $e_{ji} < \varepsilon$. The threshold value ε is called the *capture threshold* for an interval Δ_c called the *capture timeframe.* We are interested in applications where the sensor's field-of-view is much larger than the robot geometry, and hence a robot can sense a target without necessarily being close enough to capture it. Once a target is captured, it becomes inactive and is removed from the set $\mathcal{T}$; thus the game terminates when $\mathcal{T} = \varnothing$. In this game, no communication between targets and sensors takes place, but the sensors may obtain position information about the targets when they enter their fields-of-view. Based on the previous discussion, the sensing-pursuit problem can be stated as follows.

PROBLEM 2.1. *Given a set $\mathcal{P}$ of N pursuers and a set $\mathcal{T}$ of M targets moving within a specified game area $\mathcal{S}$, find a set of policies $u_p^i = c^i(q_i, Z^t) \in \mathcal{U}\ \forall i \in I_\mathcal{P}$ which maximizes the total sensing reward and minimizes the total time required to capture targets in $\mathcal{T}$ that have been positively detected.*

To complete the formulation of Problem 2.1, we define the sensing reward in terms of the probability of detection, as explained in section 3. Also, sensors and targets are modeled as hybrid systems consisting of continuous dynamics along with several discrete states [19]. Figure 2.1 shows a hierarchical state diagram for the various modes of operation. Sensors operate in one of two modes, *detection* or *pursuit*, depending on whether their primary objective is to detect targets or to capture them. Also, we assume that sensors have sufficient processing capabilities to determine the time and position of a detection event from their raw measurements.

In this problem, target tracks are classified based on the following definitions.

DEFINITION 2.2. *An unobserved track is the path of a target j for which there are no detections at the present time, t; thus $Z_j^t = \emptyset$.*

DEFINITION 2.3. *A partially observed track is the path of a target that is estimated from $1 < l < k$ individual sensor detections obtained up to the present time, t; i.e., $Z_j^t = \{z_j(t_1), \ldots, z_j(t_l)\}$.*

DEFINITION 2.4. *A fully observed track is the path of a target that is estimated from at least $k > 2$ individual sensor detections obtained up to the present time, t; i.e., $Z_j^t = \{z_j(t_1), \ldots, z_j(t_m)\}$, where $m \geq k$.*

The parameter k is chosen by the user based on the reliability of the sensor detections and on the cost associated with deploying a pursuer to capture the target. For instance, in [50] it was found that, from a geometric point of view, $k = 3$ is a convenient number of detections for estimating a track in the absence of false alarms. However, in certain surveillance applications the cost associated with capturing a target is very high, and therefore a higher number of detections may be required. Only after a track is fully observed is the target considered to be *positively detected.* Then, the estimated track is used by an upper-level controller to decide which pursuer to deploy and switch to pursuit mode, and to compute a pursuit strategy online (as described in section 3.3) that takes into account the kinematic constraints of the mobile sensors. The time that elapses between sensor i becoming a pursuer and

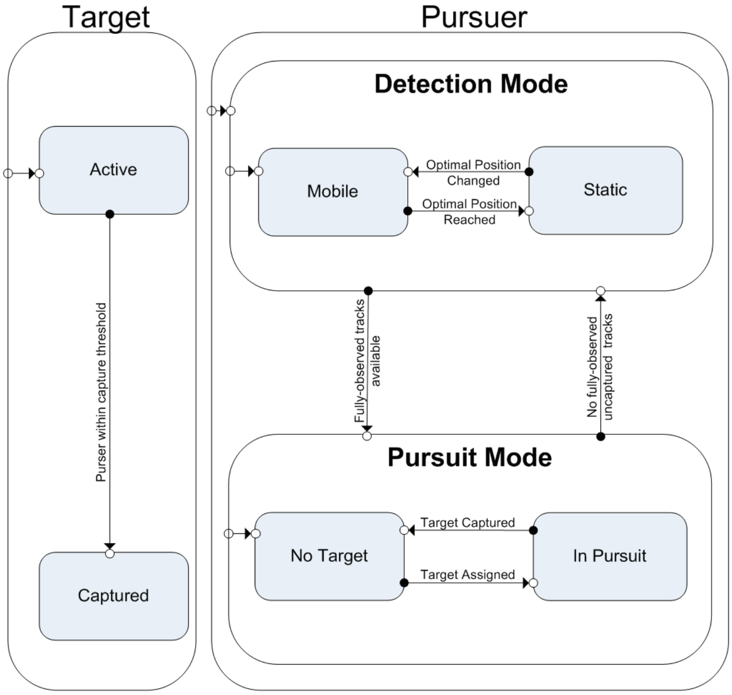


FIG. 2.1. *A finite state diagram which models the sensor (pursuer) and target as hierarchical hybrid systems with various discrete states of operation.*

intercepting target j is called the *capture time* and is denoted by $t_{c_i^j}$.

The objectives of the sensors in detection mode are to (i) avoid obstacles, (ii) maximize the probability of cooperatively detecting unobserved tracks, and (iii) maximize the probability of detecting p partially observed tracks $\{\mathcal{R}_\theta^1, \dots, \mathcal{R}_\theta^p\}$. The objectives of a sensor i in pursuit mode are to (1) avoid obstacles and (2) minimize the time $t_{c_i^j}$ required to capture a positively detected target j, based on its fully observed track $\mathcal{R}_\theta^j$ and the pursuer's position at the time of deployment. Since in practice robotic sensors are subject to kinematic (e.g., nonholonomic), dynamic, and input constraints, we have designed and implemented a simple yet effective pursuit strategy that considers the nonholonomic constraints of the mobile sensor agents used in the simulations reported in section 5.

The following section describes a methodology for planning the motions of the pursuers, in order to meet all of the above objectives.

3. Methodology. The methodology described in this section computes policies for pursuers in detection or pursuit mode that must meet multiple sensing and motion objectives. At the onset of the game, all N pursuers are placed simultaneously into $\mathcal{S}$ in detection mode. A new game round is initiated when a new partially observed or fully observed track is obtained from the latest measurement set Z^t. At every new round of the game, one pursuer in $\mathcal{P}$ is deployed and, possibly, switched to either detection or pursuit mode. Since the pursuers can perform measurements only within their fields-of-view and are installed on robotic platforms, the problems of planning the sensor measurements and the platform paths are inevitably coupled.

The primary purpose for planning the motion of the pursuers in detection mode is to obtain measurements from the targets. However, since the target tracks may be unknown (unobserved) or uncertain (partially observed), the pursuers' motion cannot be planned using classical motion planning objectives, such as minimizing distance and reaching a final configuration [29]. In fact, the positions and fields-of-view of all pursuers must be taken into account to plan the motion of a robotic sensor in a cooperative network. Thus, at every round, a pursuer's trajectory is computed based on cooperative sensing *or* pursuit objectives and, subsequently, implemented by a trajectory tracking controller that is designed based on the unicycle model (2.1).

The sensor trajectory is obtained by modifying the classical motion planning approach known as cell decomposition [29]. Let $\mathcal{C}_{free}$ denote the robotic sensors' configuration space that is free of obstacles. A cell is defined as a closed and bounded subset of $\mathcal{C}_{free}$ within which a robotic sensor path can be easily generated, and is classified based on the following properties.

DEFINITION 3.1. *A void cell is a convex polygon $\kappa \subset \mathcal{C}_{free}$ with the property that for every configuration $q_i \in \kappa$ the sensor i has zero probability of detecting a partially observed target.*

In order to account for the geometries and dynamics of the pursuers and the targets, we also introduce the following definition.

DEFINITION 3.2. *An observation cell is a convex polygon $\underline{\kappa} \subset \mathcal{C}_{free}$ with the property that for every configuration $q_i \in \kappa$ the sensor i has a nonzero probability of detecting a partially observed target.*

Void and observation cells are determined such that an obstacle-free pursuer path can be easily computed between any two configurations inside each cell. Furthermore, two cells are said to be *adjacent* if they share a common boundary and, therefore, the pursuer can move between them without colliding with the obstacles. Typically, all cells are computed such that they do not overlap. In section 3.2, a method for obtaining these cells for the system in Problem 2.1 is presented. Subsequently, they are used to obtain the following graph.

DEFINITION 3.3. *A connectivity graph, $\mathcal{G}$, is an undirected graph where the nodes represent either an observation cell or a void cell, and two nodes in $\mathcal{G}$ are connected by an arc if and only if the corresponding cells are adjacent.*

The purpose of deploying pursuers in detection mode is to detect unobserved and partially observed target tracks. Thus, the sensing objectives are expressed in terms of a reward function that represents the improvement in the overall probability of detection that would be obtained by moving from a configuration $q_i \in \kappa_l$ to a configuration in an adjacent cell, $q_i \in \kappa_\iota$,

$$R(\kappa_l, \kappa_\iota) = P_{\mathcal{R}}(\kappa_\iota) + \Delta P_{\mathcal{S}}^{k}(\kappa_l, \kappa_\iota), \tag{3.1}$$

where $\Delta P_{\mathcal{S}}^{k}$ is the gain in the probability of cooperatively detecting unobserved tracks and $P_{\mathcal{R}}$ is the probability of detecting a target with a partially observed track. These probability density functions are obtained using the methodology described in sections 3.1 and 3.2, respectively.

At the onset of the game, $Z^0 = \varnothing$, and all targets are unobserved, with $P_{\mathcal{R}} = 0$ for any cell in $\mathcal{G}$. Thus, all N pursuers in $\mathcal{P}$ are placed simultaneously in $\mathcal{S}$ by maximizing their probability of cooperatively detecting unobserved tracks at least twice, i.e., $k = 2$, such that they may be declared partially observed. Since the sensors are omnidirectional, the orientation does not influence the region covered by each field-of-view, and the pursuers are placed by determining their initial positions

$\mathcal{X}_0 = \{p_1(t_0), \dots, p_N(t_0)\}$ from the following optimization problem:

$$\mathcal{X}_0^* = \arg\max_{\mathcal{X}} \mathcal{P}_{\mathcal{S}}^2(\mathcal{X}) \tag{3.2}$$

with $0 \leq x_p^i \leq L$ and $0 \leq y_p^i \leq L$ $\forall i \in I_{\mathcal{P}}$. The above optimization amounts to a nonlinear program that can be solved by sequential quadratic programming [7, 5]. After all sensors are placed at $\mathcal{X}_0^*$, with some orientation θ_p^i, their initial configurations, $q_1(t_0), \dots, q_N(t_0)$, are known and the game begins. At every game round, a pursuer is deployed in detection or pursuit mode, depending on whether the new track is partially or fully observed, respectively. If the pursuer is switched to and deployed in pursuit mode, then its obstacle-free trajectory is computed by the method in section 3.3. If the pursuer is deployed in detection mode, its obstacle-free trajectory is computed from the sequence of cells, or *channel*, that maximizes its total reward, i.e.,

$$\mu^* \equiv \{\kappa_0, \dots, \kappa_f\}^* = \arg\max_{\mu} \sum_{(\kappa_l, \kappa_\iota) \in \mu} R(\kappa_l, \kappa_\iota), \tag{3.3}$$

where κ_f is chosen as the observation cell with the highest cumulative probability in $\mathcal{G}$, i.e., $\kappa_f = \arg\max_{\underline{\kappa}_i} (P_{\mathcal{R}}(\underline{\kappa}_i) + P_{\mathcal{S}}^k(\underline{\kappa}_i))$. In order to efficiently compute the optimal channel, μ^*, the value of the reward function (3.1) is attached to every arc in $\mathcal{G}$. Since the detection probabilities may vary slightly within each cell, they are computed in reference to the geometric centroid $\bar{q}_i$ of every cell κ_i. Then, the optimal channel μ^* is computed from $\mathcal{G}$ using the A* graph searching algorithm [29], and it is mapped into a set of waypoints that are used by a trajectory generator and trajectory tracking controller to determine the pursuer policy $u_p^i = c^i(q_i, Z^t)$.

If all sensors have the same geometry, the same connectivity-graph structure (i.e., the nodes and arcs of $\mathcal{G}$) can be utilized for all sensors. Otherwise, a different connectivity graph $\mathcal{G}_i$ may be employed for each geometry $\mathcal{A}_i$. At every round, the arc labels and the initial and final cells, κ_0 and κ_f, vary based on the latest measurements and on the sensor that is being deployed. Therefore, the A* algorithm must be run at every round of the game (section 4). In the following subsections, the probabilities of detection of unobserved and partially observed tracks that are used to define the reward function (3.1) are derived using a geometric approach.

3.1. Probability of detection for unobserved tracks.

As shown in the previous section, the observation cells in the connectivity graph represent subsets of configurations that enable measurements from partially observed tracks. Also, at any given time, the network of pursuers must detect unobserved tracks of targets that have just entered the search area in $\mathcal{S}$ that have been previously missed. Since the targets are always in motion, maximizing area coverage or other coverage formulations may not lead to effective cooperative detections. It was recently shown in [4, 49] that the quality of service of an omnidirectional sensor network performing cooperative detections of moving targets, referred to as track coverage, can be assessed without any prior knowledge of the target tracks, and depends only on the geometry of the sensors and of the search area.

In this section, track coverage is formulated using geometric transversal theory (see [23] for a comprehensive review).

DEFINITION 3.4. *A family of k convex sets in $\mathbb{R}^c$ is said to have a d-*transversal *if it is intersected by a common d-dimensional flat (or translate of a linear subspace).*

When $d = 1$ and $c = 2$, the transversal is said to be a *line stabber* of the family of convex sets. Therefore, a track detected by k sensors is a stabber of their fields-of-view, e.g., of a family $\{\mathcal{D}_1, \dots, \mathcal{D}_k\}$ in $\mathbb{R}^2$. It can be shown [17] that the family of stabbers with y-intercept b_y of a disk $\mathcal{D}_i(p_i, r_i)$ in $\mathbb{R}^2$ can be represented by the cone generated by the unit vectors

$$\hat{h}_i(b_y) = \begin{bmatrix} \cos\alpha_i & -\sin\alpha_i \\ \sin\alpha_i & \cos\alpha_i \end{bmatrix} \frac{v_i}{\|v_i\|} = \mathrm{Q}_i^+ \hat{v}_i \tag{3.4}$$

and

$$\hat{l}_i(b_y) = \begin{bmatrix} \cos\alpha_i & \sin\alpha_i \\ -\sin\alpha_i & \cos\alpha_i \end{bmatrix} \frac{v_i}{\|v_i\|} = \mathrm{Q}_i^- \hat{v}_i, \tag{3.5}$$

where $v_i \equiv p_i - [0 \quad b_y]^T$. This cone, denoted by $K(\mathcal{D}_i, b_y) = cone(\hat{l}_i, \hat{h}_i)$, is referred to as the *coverage cone* of $\mathcal{D}_i$, with origin b_y. For notational simplicity, we omit the dependency on b_y and write the above unit vectors as $\hat{l}_i = \hat{l}_i(b_y)$ and $\hat{h}_i = \hat{h}_i(b_y)$. The angle α_i denotes half the opening angle of $K(\mathcal{D}_i, b_y)$, and its sine function can be computed from the sensor position p_i:

$$\sin\alpha_i = \frac{r_i}{\|v_i\|} = \frac{r_i}{\sqrt{(x_p^i)^2 + (y_p^i - b_y)^2}}. \tag{3.6}$$

The above unit vectors are also used to determine the line stabbers of families of k nontranslate disks. We order all unit vectors in $\mathbb{R}^2$ based on the orientation of the frame $\mathcal{F}_\mathcal{S}$ (i.e., counterclockwise). Two vectors $u_i, u_j \in \mathbb{R}^2$ are said to be ordered according to the orientation of a reference frame $\mathcal{F}_\mathcal{S}$ as $u_i \prec u_j$ if, when these vectors are translated such that their origins coincide and u_i is rotated through the smallest possible angle to meet u_j, this orientation is in the same direction as the orientation of $\mathcal{F}_\mathcal{S}$ [15]. Then, the family of stabbers with y-intercept b_y can be obtained for a family of disks, as shown by the following result.

Proposition 3.5. *The set of all stabbers of a family of disks $D_k = \{\mathcal{D}_1, \dots, \mathcal{D}_k\}$, through b_y, is contained by the finitely generated cone*

$$K_k(D_k, b_y) = cone(\hat{l}^*, \hat{h}^*), \tag{3.7}$$

where

$$(\hat{l}^*, \hat{h}^*) = (\hat{l}_\imath, \hat{h}_\jmath), \quad where\, \hat{l}_\imath \succeq \hat{l}_i,\ \hat{h}_\jmath \preceq \hat{h}_j,\ \hat{l}_\imath \prec \hat{h}_\jmath, \quad \forall i, j \in I_{D_k} \tag{3.8}$$

and I_{D_k} denotes the index set of D_k. If $\hat{l}_\imath \succeq \hat{h}_\jmath$, then $K_k(D_k, b_y) = \varnothing$.

A proof is provided in Appendix A. Since $K_k(D_k, b_y)$ represents the set of tracks detected by a family of k sensors, it is referred to as the *k-coverage cone.* The opening angle of this k-coverage cone obtained by the cross product,

$$\psi = \sin^{-1} ||\hat{l}^* \times \hat{h}^*|| = H(\det[\hat{l}^* \quad \hat{h}^*]^T) \sin^{-1}(\det[\hat{l}^* \quad \hat{h}^*]^T), \tag{3.9}$$

is a Lebesgue measure over the set of line stabbers of D and is used below to obtain the probability of detection of unobserved tracks. The Heaviside function $H(\cdot)$ guarantees that if $\hat{l}^* \succ \hat{h}^*$, the opening angle of the coverage cone is equal to zero.

We restrict our attention to tracks that traverse $\mathcal{S}$ and thus intersect two of its sides. Place $\mathcal{F}_\mathcal{S}$ along two sides of $\mathcal{S}$, and a second reference frame, $\mathcal{F}'_\mathcal{S}$, along the remaining sides, as shown in Figure 3.1. Since both frames have the same orientation, Proposition 3.5 can be applied to stabbers with any intercept, namely b_y, b_x, $b_{y'}$, and $b_{x'}$ (see Figure 3.1). The opening angles of the corresponding k-coverage cones are

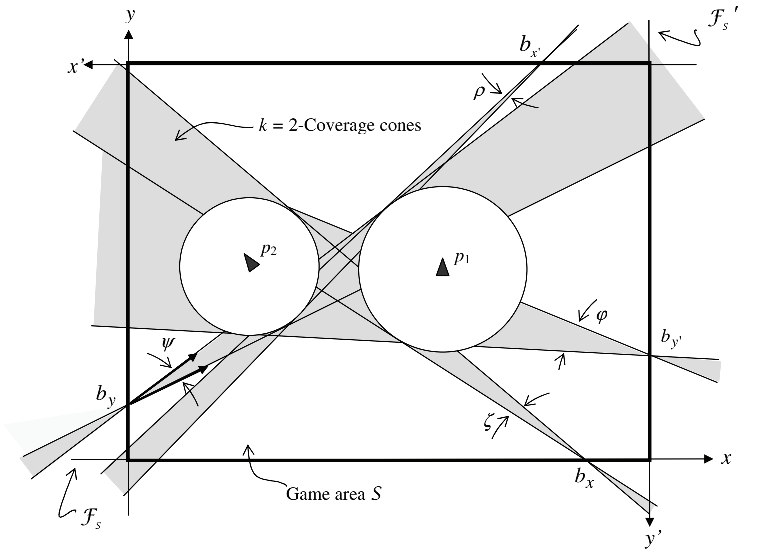


FIG. 3.1. *Coverage cone definition illustrated for two sensors with fields-of-view centered at* p_1 *and* p_2*, and a rectangular area-of-interest* $\mathcal{S}$*, with perimeter* $\partial\mathcal{S}$ *shown in bold.*

denoted by ψ, ζ, φ, and ρ, respectively, and are illustrated in Figure 3.1 for a family of $N = 2$ sensors and $k = 2$ required detections. We assume that prior to obtaining detections in $\mathcal{S}$, the probability that a target enters $\mathcal{S}$ through any intercept $b \in \partial\mathcal{S}$ and with a heading $\theta_\mathcal{T} \in (-\pi/2, \ +\pi/2)$ is uniformly distributed over all of their possible values. The set of tracks traversing $\mathcal{S}$ and intersecting at least k disks is approximated by the union of the k-coverage cones over a set of intercept values that are obtained by discretizing $\partial\mathcal{S}$ using a constant interval δb. Then, the following result can be obtained for a family of N disks representing the fields-of-view of the sensor network.

THEOREM 3.6. *The probability of detection of unobserved tracks for a set* $\mathcal{P}$ *of* N *pursuers with fields-of-view* $\mathcal{D}_1, \dots, \mathcal{D}_N$*, in a square game area* $\mathcal{S}$ *of dimensions* $L \times L$*, is a multivariate probability density function of the sensors' positions* $\mathcal{X} = \{p_1, \dots, p_N\}$ *given by a Lebesgue measure on this union,*

$$\begin{aligned} P_\mathcal{S}^k(\mathcal{X}) = {}& \frac{\delta b}{4\pi L} \sum_{\ell=1}^{L/\delta b} \sum_{j=1}^{m} (-1)^{j+1} \sum_{1 \le i_1 < \cdots < i_j \le m} [\psi(D_p{}^{i_{1,j}}, b_y^\ell) + \varphi(D_p{}^{i_{1,j}}, b_{y'}^\ell)] \\ & + \frac{\delta b}{4\pi L} \sum_{\ell=0}^{(L/\delta b - 1)} \sum_{j=1}^{m} (-1)^{j+1} \sum_{1 \le i_1 < \cdots < i_j \le m} [\zeta(D_p{}^{i_{1,j}}, b_x^\ell) + \rho(D_p{}^{i_{1,j}}, b_{x'}^\ell)] \end{aligned}$$

$$\text{with} \quad m = \frac{N!}{(N-k)!k!}, \qquad D_p{}^{i_{1,j}} \equiv \{D_k^{i_1} \cup \cdots \cup D_k^{i_j}\}, \tag{3.10}$$

where the summation $\sum_{1 \le i_1 < \cdots < i_j \le m}$ *is a sum over all the* $[m!/(m-j)!\, j!]$ *distinct integer* j*-tuples* $(i_1, \dots, i_j)$ *satisfying* $1 \le i_1 < \cdots < i_j \le m$, $D_k^{i_l}$ *denotes the* i_l*th* k*-subset of* D*, and* $D_p{}^{i_{1,j}}$ *is a* p*-subset of* D*, with* $k \le p \le n$*.*

A proof of this theorem is provided in Appendix B.

By letting $\delta b \to 0$, the Lebesgue measure (3.10) approaches the measure over

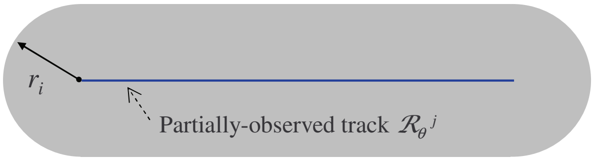


FIG. 3.2. *C-target grown isotropically from a partially observed track $\mathcal{R}_\theta^j$, based on the ith sensor range r_i.*

the entire set of tracks that traverse $\mathcal{S}$ [4]. In practice, the value δb is chosen by the user based on a trade-off between accuracy and computation time. Also, if a sensor moves to a cell κ_l, the new network configuration is approximated by $\mathcal{X}_l = \{p_1, \ldots, p_i \subset \bar{q}_l, \ldots, p_N\}$, letting the center of the sensors' field-of-view, p_i, coincide with the centroid $\bar{q}_l$ of κ_l. Thus, the gain in probability of detection for unobserved tracks that is associated with moving between two nodes $\kappa_l \rightarrow \kappa_\imath$ in $\mathcal{G}$ is

$$\Delta P^k_{\mathcal{S}}(\kappa_l, \kappa_\imath) \equiv P^k_{\mathcal{S}}(\mathcal{X}_\imath) - P^k_{\mathcal{S}}(\mathcal{X}_l). \tag{3.11}$$

The gain $\Delta P^k_{\mathcal{S}}$ is negative when the above change in configuration leads to a decreased probability of detection of unobserved tracks. However, since sensors in detection mode are moved according to (3.3) and both $P_{\mathcal{R}}$ and $P^k_{\mathcal{S}}$ pertain to the same set of targets $\mathcal{T}$, the overall probability of detection (3.1) increases at every round of the game.

3.2. Probability of detection for partially observed tracks. The partially observed tracks are viewed as an opportunity for obtaining additional measurements before investing in the costly resources needed to capture a target. In order to account for the geometry of the sensor field-of-view $\mathcal{D}_i$, the platform $\mathcal{A}_i$, and the target track $\mathcal{R}^j_\theta$, we present an approach motivated by cell-decomposition algorithms [29]. The simple philosophy behind this approach is that, in sensor planning problems, targets can be viewed as the dual of obstacles in classic robot motion planning. While in classic robot motion planning the geometry of the robot must avoid intersecting that of any obstacle, in sensor planning the geometry of the sensor's field-of-view must intersect that of the targets in order to enable sensor measurements.

Let $\mathcal{F}_{\mathcal{A}_i}$ denote a moving Cartesian frame embedded in $\mathcal{A}_i$. The configuration q_i specifies the position and orientation of $\mathcal{F}_{\mathcal{A}_i}$ with respect to the inertial frame $\mathcal{F}_{\mathcal{S}}$. If we assume that $\mathcal{D}_i$ and $\mathcal{A}_i$ are both rigid, then q_i also specifies the position of every point in $\mathcal{D}_i$ (or $\mathcal{A}_i$) relative to $\mathcal{F}_{\mathcal{S}}$. Using the latest estimate of a partially observed track, it is possible to identify the subset of $\mathcal{S}$ in which the sensors may obtain target measurements.

DEFINITION 3.7 (C-target). *The target track $\mathcal{R}^j_\theta$ in $\mathcal{S}$ maps in the ith sensor configuration space $\mathcal{C}$ to the C-target region $\mathcal{CR}_j = \{q_i \in \mathcal{C} \mid \mathcal{D}_i \cap \mathcal{R}_j \neq \varnothing,\ i \in I_{\mathcal{P}},\ j \in I_{\mathcal{T}}\}$.*

The boundary of a C-target is the curve followed by the origin of $\mathcal{F}_{\mathcal{A}_i}$ when $\mathcal{D}_i$ slides in contact with the boundary of $\mathcal{R}^j_\theta$. With the assumed robot and sensor geometries, the C-target boundaries are obtained by growing $\mathcal{R}^j_\theta$ isotropically by the radius r_i within $\mathcal{S}$, and they have the pill shape shown in Figure 3.2. C-obstacles are similarly defined [29] and are used together with the C-targets introduced above to obtain the connectivity graph $\mathcal{G}$ at every round.

Let $\mathcal{CO}_k$ denote the C-obstacle obtained from the kth obstacle in the game area,

$\mathcal{O}_k \subset \mathcal{S}$. In obstacle-avoidance algorithms, the obstacle-free configuration space,

$$\mathcal{C}^i_{free} = \mathcal{C} \setminus \bigcup_{k=1}^{n} \mathcal{CO}_k = \left\{ q_i \in \mathcal{C} \mid \mathcal{A}_i(q_i) \cap \left(\bigcup_{k=1}^{n} \mathcal{O}_k \right) = \varnothing \right\}, \tag{3.12}$$

is decomposed into a finite set of cells, $\{\kappa_1, \ldots, \kappa_f\}$, within which a path free of obstacles can be easily generated. In order to obtain a decomposition that includes observation cells (Definition 3.2), we present the following method:

(I) Decompose the configuration space that is void of any C-obstacles or C-targets and is defined as

$$\begin{aligned} \mathcal{C}^i_{void} &= \mathcal{C} \setminus \left\{ \bigcup_{j=1}^{n} \mathcal{CO}_k \;\cap\; \bigcup_{i=1}^{p} \mathcal{CR}_j \right\} \\ &= \left\{ q_i \in \mathcal{C} \mid \mathcal{A}_i(q_i) \cap \left(\bigcup_{k=1}^{n} \mathcal{O}_k \right) = \varnothing,\; \mathcal{D}_i(q_i) \cap \left(\bigcup_{j=1}^{p} \mathcal{R}_j \right) = \varnothing \right\}. \end{aligned} \tag{3.13}$$

(II) Decompose each obstacle-free C-target,

$$\mathcal{CR}_j \setminus \bigcup_{j=1}^{n} \mathcal{CO}_k, \qquad j = 1, \ldots, p, \tag{3.14}$$

thereby obtaining the set of observation cells.

(III) Construct a connectivity graph $\mathcal{G}$ using the void and observation cells obtained in (I) and (II), respectively.

When the C-targets are grown isotropically by a disk (Figure 3.2), the decomposition may involve generalized polygons [29]. A sweeping-line algorithm can be used to decompose a nonconvex generalized polygon with ν vertices into $O(\nu)$ convex generalized polygons in $O(\nu \log \nu)$ time (see section 5.1 in [29]). Alternatively, the pill-shaped C-targets can be approximated by a convex polygon, obtaining the running time presented in section 4. An illustrative example of workspace and corresponding cell decomposition is shown in Figure 3.3. The connectivity graph constructed using this cell decomposition is illustrated in Figure 3.4, where the observation cells are shown in grey and the void cells are white. Each node in the connectivity graph corresponds to one polygonal cell in Figure 3.3, where the cells are numbered from left to right and from top to bottom.

The probability density function $P_{\mathcal{R}}$, used to compute the reward (3.1), is obtained as follows. Suppose that $\underline{\kappa}_l$ is one of the observation cells that are obtained from the decomposition of the jth C-target: $\underline{\kappa}_l \subset \mathcal{CR}_j$. Then, the sensing benefit of visiting the lth cell in $\mathcal{G}$ is the probability of detecting the target j,

$$P_{\mathcal{R}}(\underline{\kappa}_l \subset \mathcal{CR}_j) = \Pr\{D_{ji} = 1 \mid e_{ji} \leq r_i\}, \tag{3.15}$$

where D_{ji} represents the event that the ith sensor reports a detection when the jth target comes within its detection range. In this paper, $P_{\mathcal{R}}$ is assumed to be uniform over $\mathcal{CR}_j$ for simplicity, and when a cell is void $P_{\mathcal{R}}(\kappa_l) = 0$ since the sensor field-of-view will not intersect any of the p partially observed tracks. In general, $P_{\mathcal{R}}$ can be estimated from knowledge of the measurement process and can be made dependent on time and on the distance from the target [18].

The next section presents an effective control methodology by which sensors in pursuit mode capture and intercept targets whose tracks have been fully observed and thus have been declared positively detected.

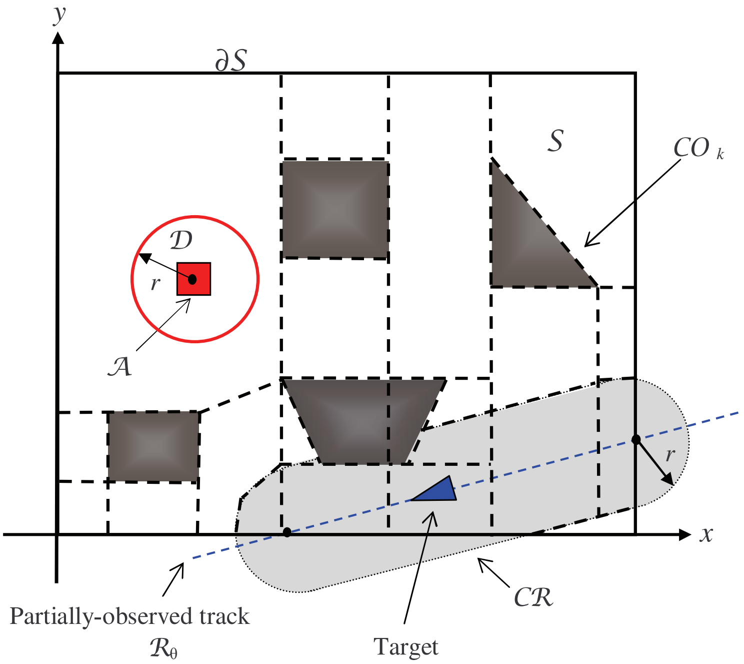


FIG. 3.3. *Example of cell decomposition (dashed lines) for a workspace with four C-obstacles (darkly shaded polygons) and one C-target $\mathcal{CR}$ (lightly shaded region) corresponding to $2 < k$ detections. One sensor with range r and field-of-view $\mathcal{D}$ is installed on a robot with a square platform geometry $\mathcal{A}$.*

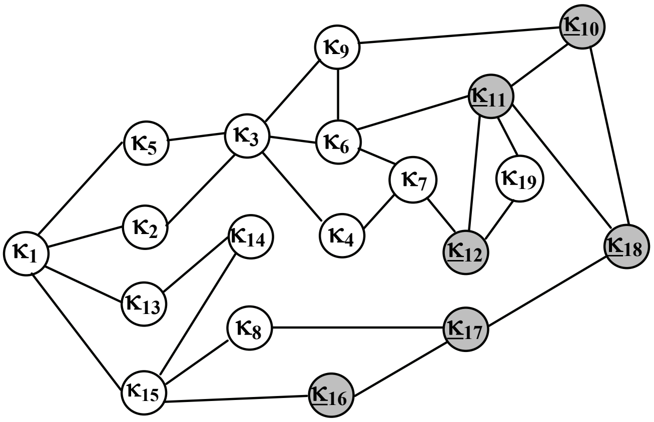


FIG. 3.4. *Connectivity graph obtained from the cell decomposition in Figure* 3.3*, where the cells in the decomposition are numbered from left to right and from top to bottom, and the observation cells are shown in grey.*

3.3. Pursuit strategy. Once a new target is positively detected, a sensor is switched to pursuit mode and deployed to capture it. A geometric approach motivated by the behavior of the potential field controller, described in [10], is used to drive the

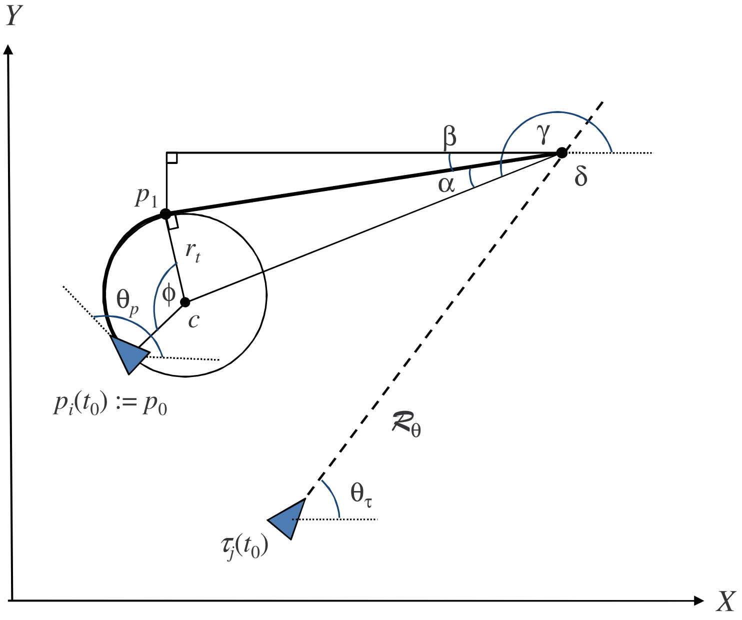


FIG. 3.5. *Control strategy to capture a target.*

nonholonomic robot sensor in pursuit mode to a goal waypoint, $\delta \in \mathbb{R}^2$, calculating both the point and time at which a pursuer will intercept a target moving in a known straight line, $\mathcal{R}^j_\theta$, with constant velocity, V_τ. This strategy, depicted in Figure 3.5, is based upon the geometry of the problem and takes into account the kinetic constraints of the pursuer but not the presence of the obstacles. Therefore, here it is combined with the cell-decomposition methodology presented in section 3.2.

First, the interception point δ is calculated by determining the time required by both the pursuer and target to reach δ. The pursuit initial time t_0 can be assumed to be the time at which the last detection $z_j(t_k)$ became available from target j, and $p_i(t_0)$ and $\tau_j(t_0)$ denote the initial positions of the pursuer i and target j, respectively. The interception point δ and the time to interception $t_{c_i^j}$ are defined as

$$(3.16) \qquad \delta = \begin{bmatrix} x^j_\tau(t_0) + t_{c_i^j} V_\tau \cos\theta^j_\tau \\ y^j_\tau(t_0) + t_{c_i^j} V_\tau \sin\theta^j_\tau \end{bmatrix}, \qquad t_{c_i^j} = \frac{r_t\phi + \|c-\delta\|\cos\alpha}{V_{p_{\max}}},$$

where the distance traveled by the pursuer is the distance along the arc p_0p_1 plus the straight line distance between p_1 and δ. The arc radius is the same as the turn radius of the pursuer and is defined as $r_t = \frac{V_{p_{\max}}}{\omega_p}$, where $V_{p_{\max}}$ and ω_p are the maximum speed and angular velocity, respectively, of the pursuer. There are two possible circles corresponding to a right or a left turn of the pursuer. The center points, c_R and c_L, of the circles defined by the turn radius are calculated as

$$c_R = \begin{bmatrix} p_{0x} + r_t\cos(\theta_p - \frac{\pi}{2}) \\ p_{0y} + r_t\sin(\theta_p - \frac{\pi}{2}) \end{bmatrix}, \qquad c_L = \begin{bmatrix} p_{0x} + r_t\cos(\theta_p + \frac{\pi}{2}) \\ p_{0y} + r_t\sin(\theta_p + \frac{\pi}{2}) \end{bmatrix}.$$

The center point lying closest to the interception point is chosen as

$$c = \begin{cases} c_R & \text{if } \|c_R-\delta\| \le \|c_L-\delta\|, \\ c_L & \text{if } \|c_R-\delta\| > \|c_L-\delta\|. \end{cases}$$

The other parameters for calculating the interception time are calculated as

$$\alpha = \arcsin\left(\frac{r_t}{\|c-\delta\|}\right), \qquad \gamma = \arctan\left[c_y - \delta_y, c_x - \delta_x\right],$$

$$\beta = \begin{cases} \gamma - \alpha & \text{if } c = c_R, \\ \gamma + \alpha & \text{if } c = c_L, \end{cases} \qquad p_1 = \begin{bmatrix} \delta_x + \|c-\delta\| \cos\alpha \cos\beta \\ \delta_y + \|c-\delta\| \cos\alpha \sin\beta \end{bmatrix},$$

$$\phi = \left|\arctan(p_{1y} - c_y, p_{1x} - c_x) - \arctan(p_{0y} - c_y, p_{0x} - c_x)\right|,$$

where $\delta = [\delta_x \ \delta_y]^T$. The time to interception $t_{c_i^j}$ and the interception point δ in (3.16) are computed numerically by Newton's method [34].

In order to find an obstacle-free shortest path between $p_i(t_0)$ and δ, the connectivity graph $\mathcal{G}$ obtained in section 3.2 is modified by changing the arc labels to reflect the Euclidean distance between any two nodes $\kappa_l \to \kappa_\imath$ in $\mathcal{G}$:

$$d(\kappa_l, \kappa_\imath) \equiv \max ||\mathcal{A}(\bar{q}_\imath) - \mathcal{A}(\bar{q}_l)||. \tag{3.17}$$

The channel μ_p^* of shortest overall distance between $\kappa_0 \ni q_i(t_0)$ and $\kappa_f \ni \delta$ (assuming a zero heading at δ) can be determined by the graph searching algorithm A* [29]. Subsequently, μ_p^* is mapped into a set of waypoints in $\mathbb{R}_+^2$ that are used by an inner-loop trajectory generator and trajectory tracking controller designed for the unicycle model (2.1).

4. Performance and complexity analysis.

Previous work on the correctness and complexity of pursuit-evasion games has focused on graphs, in which one or more pursuers attempt to capture one target by moving between adjacent nodes in a graph (see [24, 2, 28, 38] for a comprehensive review). In these problem formulations, the sensing ability and fields-of-view of the pursuers are not taken into account, and the pursuit strategies consist of randomized searches on the graph, because the pursuers cannot see the evader until the latter is caught. Also, only one evader who may be restricted or unrestricted to the graph is considered during each game. By computing the connectivity graph by the methodology in section 3.2, these results could potentially be extended to the pursuit-evasion game in Problem 2.1. For example, if the strategy in [2] is implemented for one pursuer and one evader ($N = M = 1$), then the pursuer captures the evader on an n-node cycle with probability at least $\Omega(1/\log(n_\mathcal{G}))$, and the game ends in $O(n_\mathcal{G} \ \log(\text{diam}(\mathcal{G})))$ time, where $n_\mathcal{G}$ is the number of nodes in $\mathcal{G}$ and $\text{diam}(\mathcal{G})$ is the diameter of the graph. However, by not taking into account the sensing ability of the pursuers and the knowledge of fully observed tracks (i.e., the presence of observation cells), these strategies are not very effective at capturing multiple evaders in large game areas. In these applications, $n_\mathcal{G}$ and $\text{diam}(\mathcal{G})$ are very large. Therefore, the probability of capturing the evaders can be very small and the game end time $O(Mn_\mathcal{G}\log(\text{diam}(\mathcal{G}))$ can be very large.

The correctness and game end time for the strategy presented in section 3 are analyzed by assuming that the time required to maneuver around obstacles or to turn (ϕ/ω_p) are negligibly small compared to the duration of the game. Let $(\bar{\cdot})$ denote the expected value (or mean), and $\lfloor\cdot\rfloor$ denote the floor function. Then, the performance of the sensor network depends on the dimension of the game area (L), the number of sensors (N), the number of required detections ($k > 2$), and the field-of-view radius

(r_i), which here is assumed constant $(r_i = r\ \forall i)$ for simplicity, as summarized by the following result.

THEOREM 4.1. *The pursuit-evasion game in Problem* 2.1 *is guaranteed to terminate, provided that*

$$N \geq N_{\min} = \frac{1}{2}\left[\left\lfloor \frac{2L}{r} \right\rfloor + k - 1 + \left|\left\lfloor \frac{2L}{r} \right\rfloor - k + 3\right|\right], \tag{4.1}$$

and requires a time

$$t_f \leq T_u = \frac{(\sqrt{2}L - 2r)}{V_{\mathcal{T}_{\min}}} + \left[\left\lfloor \frac{(k-2)M}{N} \right\rfloor + 1\right] \frac{(\sqrt{2}L - r)}{\bar{V}_p} \tag{4.2}$$

$$+ \frac{r}{(V_{p_{\max}}^2 - \bar{V}_{\mathcal{T}}^2)} + \left\lfloor \frac{M}{N} \right\rfloor \frac{\left(\bar{V}_{\mathcal{T}} + \sqrt{2V_{p_{\max}}^2 - \bar{V}_{\mathcal{T}}^2}\right)}{(V_{p_{\max}}^2 - \bar{V}_{\mathcal{T}}^2)^2} L \tag{4.3}$$

to capture all M targets in $\mathcal{T}$. If the network contains at least

$$N_\ell = \frac{1}{2}\left[\ell \left\lfloor \frac{2L}{r} \right\rfloor - 4\ell(\ell - 1) + (k-2)M + \left|\ell \left\lfloor \frac{2L}{r} \right\rfloor - 4\ell(\ell - 1) - (k-2)M\right|\right] \tag{4.4}$$

sensors, with $\ell = 1, \dots, \lfloor L/4r \rfloor$, then all targets in $\mathcal{T}$ can be captured in a time

$$\begin{aligned} t_f \leq T_\ell = {} & \frac{1}{V_{\mathcal{T}_{\min}}}\left\{\frac{\sqrt{2}}{2}L - 2\sqrt{2}r(\ell - 1) + \left|2r[1 + \sqrt{2}(\ell - 1)] - \frac{\sqrt{2}}{2}L\right|\right\} \\ & + \frac{(\sqrt{2}L - r)}{\bar{V}_p} + \frac{r}{(V_{p_{\max}} - \bar{V}_{\mathcal{T}})}, \end{aligned} \tag{4.5}$$

and the game terminates in $t_f \leq T_\ell \leq T_u$, where $T_\ell = T_u$ when $\ell = 1$ and $k = 3$.

A proof is provided in Appendix C. Based on the above result, $N_{\min}$ is the minimum number of sensors needed to guarantee that the game will end in less than T_u time. But, if more sensors can be utilized, then N can be increased according to (4.4) to decrease the maximum time required to end the game, as shown by (4.5). The performance of the algorithm also depends on the choice of reward function (3.1) through the parameter k, which, together with the parameters N, L, and r, specifies the definition of the probability density function (3.10).

In Problem 2.1, a round is defined as the deployment of one sensor in either detection or pursuit mode, and it is initiated based on the measurement set Z^t when a new target track becomes either partially observed or fully observed. Thus, the computational complexity of the algorithm in section 3 is assessed based on the calculations required by each round. Let $n_{e_{\mathcal{O}}}$ denote the number of edges required to describe all n obstacles in $\mathcal{S}$, and $n_{e_{\mathcal{R}}} = 2p$ denote the number of edges required to describe all p tracks that have been partially observed up to the present time. Then, if $n_{\underline{\mathcal{K}}}$ and $n_{\mathcal{G}_a}$ are the number of observation cells and the number of arcs in $\mathcal{G}$, respectively, and $n_\delta \equiv L/\delta b$, the following result is obtained.

THEOREM 4.2. *In every round of the pursuit-evasion game in Problem* 2.1, *the running time required to deploy a sensor in detection mode is*

$$\begin{aligned} \Gamma_d = {} & O((n_{e_{\mathcal{O}}} + n_{e_{\mathcal{R}}}) \log(n_{e_{\mathcal{O}}} + n_{e_{\mathcal{R}}}) + n_{e_{\mathcal{R}}} n_{e_{\mathcal{O}}} \log n_{e_{\mathcal{O}}}) + O(n_{\underline{\mathcal{K}}} n_\delta m(k + \log m)) \\ & + O(n_{\mathcal{G}}^2 + n_{\mathcal{G}_a}), \end{aligned} \tag{4.6}$$

where $m = \binom{N}{k}$ *is given by the binomial coefficient and the running time required to deploy a sensor in pursuit mode is*

$$\Gamma_p = O((n_{\mathcal{G}} + n_{\mathcal{G}_a}) \log_2 n_{\mathcal{G}}). \tag{4.7}$$

A proof is provided in Appendix D. Clearly, depending on the characteristics of the robotic sensors and of the workspace, $\mathcal{S}$, only one of the three terms in (4.6) will dominate over the others, providing the overall running time complexity of the detection round. As an example, when the leading time complexity is that of the reward function (3.1), the deployment of a sensor in detection mode in a problem with $N = 50$, $k = 3$, $r = 5$ km, and $L = 100$ km took 0.797 sec on a Pentium 4 CPU 3.06 GHz computer. On the same computer, when the leading time complexity is that of A*, the deployment of a sensor in pursuit mode took between 0.078 and 0.2350 sec in a connectivity graph with $n_{\mathcal{G}} = 340$ and $n_{\mathcal{G}_a} = 338$, and between 1.672 and 60.125 sec in a graph with $n_{\mathcal{G}} = 9{,}590$ and $n_{\mathcal{G}_a} = 32{,}687$.

5. Simulation results. In order to validate the methodology developed in section 3, a MATLAB simulator has been developed. We integrate the information-driven sensor planning and pursuit strategies described in previous sections into several simulation scenarios.

5.1. Scenario 1: Multiple static sensors and one pursuer. In many surveillance applications static sensor networks can be used with a few motion-enabled sensors. Static sensors are placed to optimally cover a given area [17]. If a target (evader) is detected, then a mobile sensor can be sent to investigate or capture the target. This scenario is a special case of the pursuit-evasion problem addressed in this work. The simulation results are depicted in Figure 5.1. This scenario includes obstacles and the use of the reward function (3.1).

5.2. Scenario 2: Multiple mobile sensors and targets. This simulation scenario extends the first by considering the same environment but with multiple targets. Before the simulation scenario begins, five sensors with platforms measuring 0.25 m square are placed in the 10m-by-10m environment to maximize the probability of detecting tracks with $k = 2$, since we require this number of detections to form a partially observed track. Obstacle and coverage maps are generated for each sensor corresponding to the placement in each cell. Figure 5.2 shows the initial environment and the five sensors—one with sensing radius 1.5 m, one with sensing radius 1.25 m, and three with sensing radii of 1 m. Initially, all sensors are in *detection* mode, and each is a candidate to switch to the *pursuit* mode when target tracks become fully observed.

In this scenario, two targets enter the environment at different locations and headings and with different velocities. As they move along their trajectories, they are detected by the sensors (Figure 5.3). The sensors remain motionless since each target has been detected only once. After the second detection of a target, the network hypothesizes the target track based on previous detections and deploys the sensor which receives the highest reward (or lowest cost) as obtained by the A* graph search algorithm to move to obtain an additional detection of the target (Figure 5.4). When the second target becomes partially detected, the same track hypothesis and sensor deployment occurs. At the point that the first target's track becomes fully observed (see Figure 5.5), the network again evaluates the reward (distance) and deploys the best sensor to pursue the target. The same pursuit is performed when the second

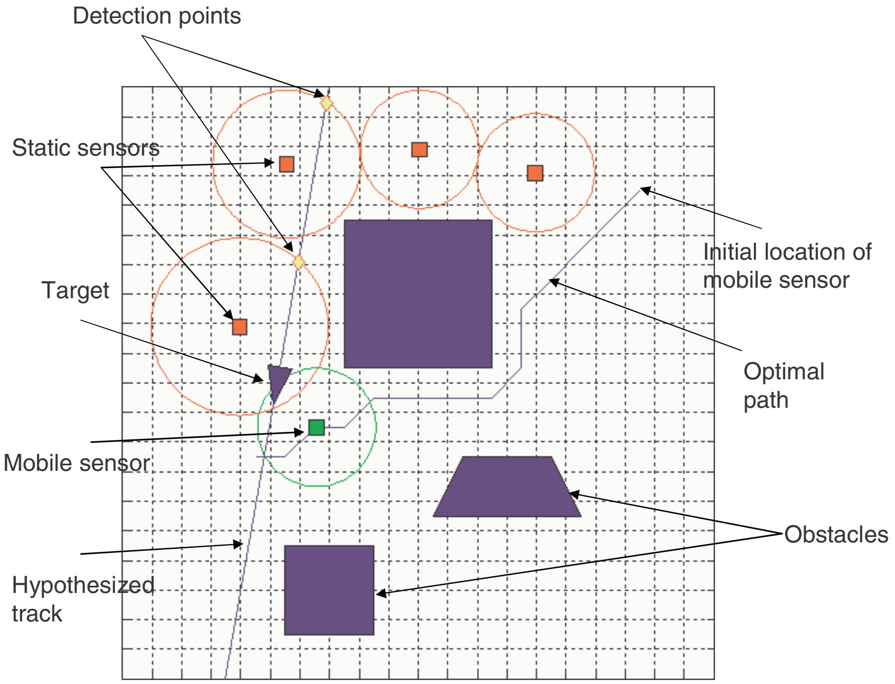
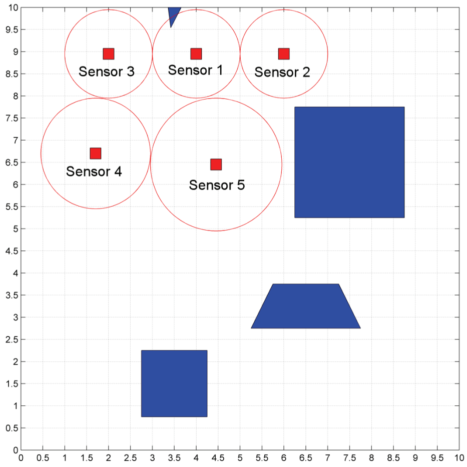


Fig. 5.1. *Static multiple detectors and one pursuer.*



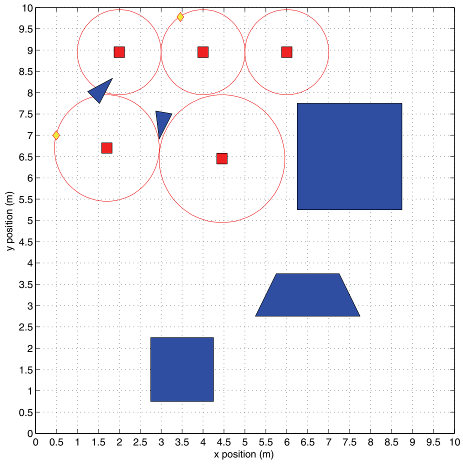


Fig. 5.2. *Initial sensor placement.*

Fig. 5.3. *Two targets each detected once.*

target is fully observed, as shown in Figure 5.6. The state of the network following capture of all known targets is depicted in Figure 5.7. The network is rearranged to maximize area coverage at the next recalculation interval. Table 5.1 summarizes the chronology of the main events which occur during the simulation. Algorithm 1 illustrates how the simulation scenario has been implemented.

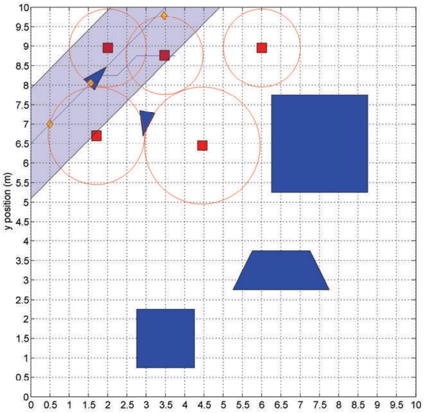


FIG. 5.4. *Target 1 is partially observed with its hypothesized track, and Sensor 1 is deployed to obtain additional observations.*

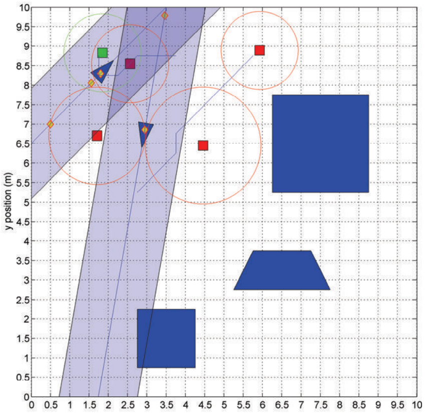


FIG. 5.5. *Target 1 is fully observed and Sensor 3 is deployed to pursue it while Target 2 becomes partially observed, and Sensor 1 is deployed to obtain an additional observation.*

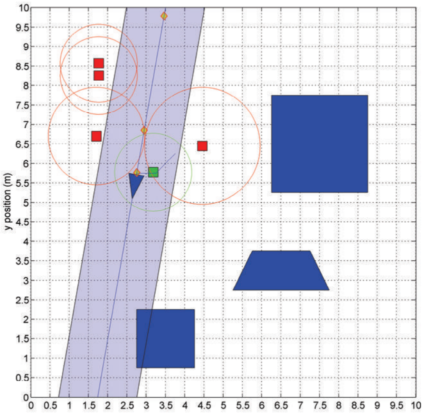


FIG. 5.6. *Target 1 has been captured. Target 2 is fully observed and is pursued by Sensor 1.*

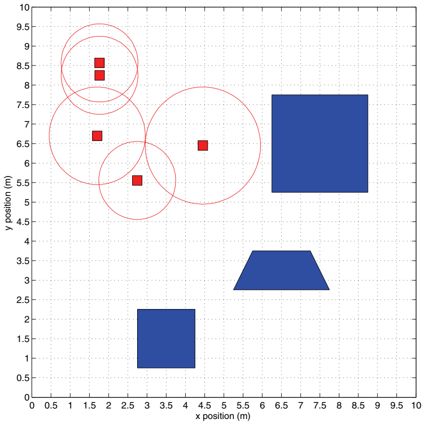


FIG. 5.7. *Final sensor arrangement after both targets are captured.*

TABLE 5.1
Simulation events of Scenario 2.

Event	Time (s)	Position (m)	Sensor	Target
Detect	0.40	(3.46,9.78)	1	2
Detect	1.70	(0.49,6.99)	4	1
Detect	5.45	(1.56,8.06)	3	1
Deploy	5.45	(1.75,8.25)	1	1
Detect	6.30	(1.80,8.30)	1	1
Pursue	6.30	-	3	1
Detect	6.35	(2.94,6.85)	4	2
Deploy	6.35	(2.75,5.25)	2	2
Capture	6.60	(1.76,8.57)	3	1
Detect	8.55	(2.75,5.77)	2	2
Pursue	8.55	-	2	2
Capture	9.40	(2.75,5.55)	2	2

Algorithm 1. Scenario 2 Algorithm.

```
Perform initial optimal sensor placement
Decompose environment into $\mathcal{C}_{free}$ and $\mathcal{C}_{obstacle}$ cells
for all Sensors do
   Calculate obstacle map
   Calculate coverage map
end for
while Game not over do
   for all Sensors in pursuit do
      if Pursued target beneath capture threshold then
         Remove target
         End pursuit
      end if
   end for
   if Detection then
      if Target detections = 2 then
         Hypothesize target track
         Calculate observation cells
         for all Sensors that have not detected this target do
            Calculate path and reward to investigate target
         end for
         Deploy the sensor with the greatest reward
      else if Target detections = 3 then
         for all Sensors not in pursuit do
            Calculate path and reward to pursue target
         end for
         Deploy the sensor with the greatest reward
      end if
   end if
   if Sensor update interval then
      for all Sensors do
         Calculate coverage map
      end for
      Deploy next sensor to maximize coverage
   end if
end while
```

6. Conclusions. This paper presents a novel framework for developing sensor control policies in systems involving multiple robotic platforms that seek to detect and intercept multiple mobile targets. Multiple objectives, such as the probability of detecting unobserved tracks, for which little or no information is available a priori, obstacle avoidance, and the probability of detection associated with partially observed targets are approached using a geometric approach. The path leading to the optimal trade-off between these objectives is obtained through the A* graph searching algorithm and is passed to a control strategy that accounts for the actual pursuers' dynamics. By adopting a track-before-detect approach, a target is declared positively detected once a satisfactory number of detections k may be used to form a consistent track. Subsequently, a heuristic rule switches one of the mobile sensors from *detection* mode to *pursuit* mode, and the track is readily available to compute an optimal pursuit strategy. By maximizing the same reward function, the remaining sensors in *detection* mode are reconfigured such that the probability of detecting the remaining targets is again optimized. The progressive simulation scenarios presented validate the developed methodology. The future work of this approach will include fully implementing the methodology on a multivehicle testbed [13]. Additionally, the approach will be extended to reflect a more general pursuit-evasion game by considering intermittent communication among pursuers and targets, intermittent estimation, and intelligent evaders that are not restricted to moving in straight lines.

Appendix A. Proof of Proposition 3.5. This proof considers a family of $k = 3$ nontranslates $D_k = \{\mathcal{D}_i, \mathcal{D}_j, \mathcal{D}_l\}$ with index set $I_{D_k} = \{i, j, l\}$. The results can be extended to higher k by induction. The coverage cone $K(\mathcal{D}_\ell, b_y)$ contains the set of all rays that intersect $\mathcal{D}_\ell$ in $\mathbb{R}^2_+$, where $\ell \in I_{D_k}$. Then, the set of tracks intersecting all circles in the family D_k is given by the following intersection:

$$K_k(D_k, b_y) = \bigcap_{\ell \in I_{D_k}} K(\mathcal{D}_\ell, b_y) = K(\mathcal{D}_i, b_y) \cap K(\mathcal{D}_j, b_y) \cap K(\mathcal{D}_l, b_y). \tag{A.1}$$

From the properties of cones [6, p. 70], the intersection of a collection of cones is also a cone, and thus $K_k(D_k, b_y)$ is a cone. A vector z representing a ray $\mathcal{R}_\theta$ with the same slope and origin lies in a cone K if and only if $\mathcal{R}_\theta$ lies in K, since any point on $\mathcal{R}_\theta$ can be written as cz, with $c > 0$.

Consider a ray $\mathcal{R}_\theta \in K(\mathcal{D}_\ell, b_y)$, where $K(\mathcal{D}_\ell, b_y) = cone(\hat{l}_\ell, \hat{h}_\ell)$ and thus can be represented by a vector $z_\ell = c_1 \hat{l}_\ell + c_2 \hat{h}_\ell$ with constants $c_1, c_2 > 0$. Then, $z_\ell \in K(\mathcal{D}_\ell, b_y)$ and, by the properties of vector sum, $\hat{l}_\ell \prec z_\ell \prec \hat{h}_\ell$. Next, consider a cone $K^* = cone(\hat{l}^*, \hat{h}^*)$ that is finitely generated by two unit vectors $\hat{h}^* = \hat{h}_\jmath$ and $\hat{l}^* = \hat{l}_\imath$ with $\jmath, \imath \in I_{D_k}$, and assume $\hat{l}_\imath \prec \hat{h}_\jmath$. By the properties of finitely generated cones [6], any vector $z^* = b_1 \hat{l}^* + b_2 \hat{h}^*$ with constants $b_1, b_2 > 0$ must lie in K^*. It follows that a ray $\mathcal{R}^*_\theta$ with the same slope and origin as z^* must also lie in K^*, since any point on $\mathcal{R}^*_\theta$ can be written as $cz*$, with $c > 0$. Since z^* is a positive combination of $\hat{l}^*$ and $\hat{h}^*$, it also follows that $\hat{l}^* \prec z^* \prec \hat{h}^*$.

According to Proposition 3.5, choose $\hat{h}^* = \hat{h}_\jmath \preceq \hat{h}_\ell$ and $\hat{l}^* = \hat{l}_\imath \succeq \hat{l}_\ell$ $\forall \ell \in I_{D_k}$. Suppose that the unit vectors of D_k can be ordered as $\hat{h}_l \prec \hat{h}_j \prec \hat{h}_i$ and $\hat{l}_i \prec \hat{l}_l \prec \hat{l}_j$. Then, the unit vectors and z^* can be ordered as follows:

$$\hat{l}_\ell \preceq \hat{l}_j = \hat{l}^* \prec z^* \prec \hat{h}^* = \hat{h}_l \preceq \hat{h}_\ell \qquad \forall \ell \in \{i, j, l\} = I_{D_k} \tag{A.2}$$

or, more explicitly,

$$\hat{l}_i \prec \hat{l}_l \prec \hat{l}_j = \hat{l}^* \prec z^* \prec \hat{h}^* = \hat{h}_l \prec \hat{h}_j \prec \hat{h}_i. \tag{A.3}$$

Since the above order also implies $\hat{l}_\ell \prec z^* \prec \hat{h}_\ell$ $\forall \ell \in I_{D_k}$, then $z^*, \mathcal{R}^*_\theta \in K(\mathcal{D}_\ell, b_y)$ $\forall \ell \in I_{D_k}$. Thus, from (A.1), $z^*, \mathcal{R}^*_\theta \in K_k(D_k, b_y) = K^* = cone(\hat{l}^*, \hat{h}^*)$, provided that $\hat{h}^*$ and $\hat{l}^*$ are chosen subject to (A.2).

So far it has been assumed that $\hat{l}_\imath \prec \hat{h}_\jmath$. If the unit vectors of D_k are such that $\hat{l}_\imath \succ \hat{h}_\jmath$, then there are no vectors that can satisfy the order $\hat{l}_\imath = \hat{l}^* \prec z^* \prec \hat{h}^* = \hat{h}_\jmath$, and $K_k(D_k, b_y) = K^* = \varnothing$.

Appendix B. Proof of Theorem 3.6. The set of all tracks through a y-intercept b_y that are detected by at least k sensors in $D = \{\mathcal{D}_1, \ldots, \mathcal{D}_N\}$ is the union of the k-coverage cones of all k-subsets of D,

$$\mathcal{K}_k(D, b_y) = \bigcup_{j=1}^{m} K_k(D_k^j, b_y), \quad m = \binom{N}{k}. \tag{B.1}$$

D_k^j denotes the jth k-subset of D, and the number m of possible k-subsets is given by the binomial coefficient N *choose* k. Since $\mathcal{K}_k(D, b_y)$ is a union of possibly disjoint cones, it may not be a cone [6]. Nevertheless, the same Lebesgue measure defined for a cone, μ on $[0,\ \pi]$, can be applied to it using the principle of inclusion-exclusion [43]

$$\begin{aligned} \mu(\mathcal{K}_k(D, b_y)) &= \mu\left(\bigcup_{j=1}^{m} K_k(D_k^j, b_y)\right) \\ &= \sum_{j=1}^{m}(-1)^{j+1} \sum_{1 \le i_1 < \cdots < i_j \le m} \mu(K_k(D_k^{i_1}, b_y) \cap \cdots \cap K_k(D_k^{i_j}, b_y)), \end{aligned} \tag{B.2}$$

where $m = \binom{N}{k} = \frac{N!}{(N-k)!\ k!}$ and $\sum_{1 \le i_1 < \cdots < i_j \le m}$ is a sum over all the $[m!/(m-j)!\ j!]$ distinct integer j-tuples $(i_1, \ldots, i_j)$ satisfying $1 \le i_1 < \cdots < i_j \le m$. Also, $\mu(\cdot)$ denotes a measure on the set. Since the right-hand side of (B.2) is an intersection of cones, it also is a cone on which we can impose μ. Moreover, it represents the set of tracks through b_y that intersect all sensors in $D_p^{i_{1,j}} = \{D_k^{i_1} \cup \cdots \cup D_k^{i_j}\}$. Based on the properties of k-subsets, $D_p^{i_{1,j}}$ must contain $k \le p \le n$ elements of D and, thus, is a p-subset of D. Based on the properties of k-coverage cones (Proposition 3.5), the set of line transversals of $D_p^{i_{1,j}}$ through b_y can be represented by the p-coverage cone $K_p(D_p^{i_{1,j}}, b_y)$. Thus, (B.2) can be written as

$$\begin{aligned} \mu(\mathcal{K}_k(D, b_y)) &= \sum_{j=1}^{m}(-1)^{j+1} \sum_{1 \le i_1 < \cdots < i_j \le m} \mu(K_p(D_p^{i_{1,j}}, b_y)) \\ &= \sum_{j=1}^{m}(-1)^{j+1} \sum_{1 \le i_1 < \cdots < i_j \le m} \psi(D_p^{i_{1,j}}, b_y), \end{aligned} \tag{B.3}$$

by Proposition 3.5, where p is the number of elements in the union of j k-subsets of D, and the opening angles $\psi(D_p^{i_{1,j}}, b_y)$ in the above summation are given by (3.9).

Now, let $\mathcal{R}_\theta(b_y)$ denote a ray with y-intercept b_y and heading angle $\theta_\tau \in (-\pi/2,\ +\pi/2)$. Suppose $\mu(\mathcal{K}_k(D, b_y)) = \pi$; then $\mathcal{R}_\theta(b_y)$ will be detected by k pursuers with probability one. Assuming that all headings are equally likely, if $0 \le \mu(\mathcal{K}_k(D, b_y)) \le \pi$, then $\mathcal{R}_\theta(b_y)$ will be detected by k pursuers with probability $\mu(\mathcal{K}_k(D, b_y))/\pi$. Let $P(b_y)$ denote the prior probability that a target enters $\mathcal{S}$ at b_y. Assuming that all y-intercepts are equally likely, $P(b_y) = \delta b/(L + \delta b)$. Since the two events are independent, the probability that an unobserved track is $\mathcal{R}_\theta^j(b_y)$ and is detected by k pursuers is given by the product of the individual probabilities

$$\Pr\{D_{jk} = 1, \mathcal{R}_\theta^j(b_y)\} = \frac{\delta b}{(L + \delta b)} \frac{\mu(\mathcal{K}_k(D, b_y))}{\pi}, \tag{B.4}$$

where D_{jk} denotes the event that the jth target traversing $\mathcal{S}$ along an unobserved track is detected by at least k pursuers. Since the k-coverage cones with different

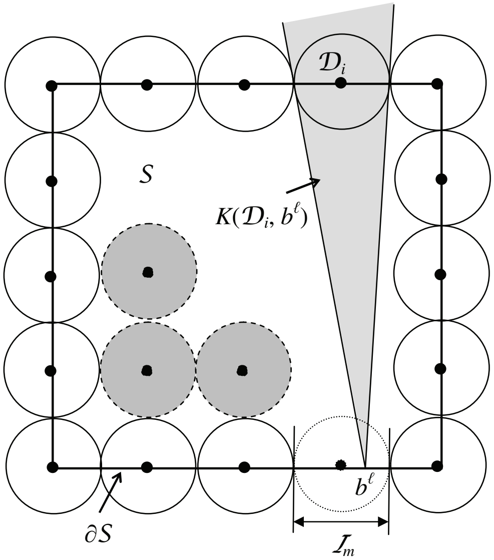


FIG. B.1. *Examples of total coverage configuration for $k = 2$ (sensors on $\partial\mathcal{S}$), sensors packed by the triangular number (grey disks), and coverage cone $K_2 < \pi$ when $N < 2L/r$.*

y-intercepts are disjoint, the probability that a target's point of entry in $\mathcal{S}$ is the ℓth intercept $b_y^\ell \in \partial\mathcal{S}$ and that it is detected by k pursuers is

(B.5)

$$\Pr\{D_{jk} = 1, \mathcal{R}_\theta^j(b_y^\ell \in \partial\mathcal{S})\} = \sum_{\ell=0}^{L/\delta b} \frac{\delta b}{\pi(L+\delta b)} \sum_{j=1}^{m} (-1)^{j+1} \sum_{1 \le i_1 < \cdots < i_j \le m} \psi(D_p^{i_1,j}, b_y^\ell).$$

The probability that $D_{jk} = 1$ and that the target's point of entry is on one of the other axes can be obtained by the same approach, using the opening angles of the corresponding coverage cones. The set of tracks that traverse $\mathcal{S}$ and are detected by at least k pursuers is given by the probability of the union of intersecting sets $\mathcal{K}_k(D, b_y)$, $\mathcal{K}_k(D, b_x)$, $\mathcal{K}_k(D, b_{y'})$, and $\mathcal{K}_k(D, b_{x'})$ that are obtained by applying (B.1) to intercepts on the y, x, y', and x' axes, respectively. The final probability density function $P_\mathcal{S}^k(\mathcal{X})$ in (3.10) is obtained by observing that every track in this union intersects two sides of $\mathcal{S}$, and that the indices in the second summation must be shifted in order to consider intercepts at the corners only once.

Consider now the case of a *total coverage configuration*, denoted by $\mathcal{X}_{tot}^k$, that detects all tracks in $\mathcal{S}$ at least k times. We want to show that for $\mathcal{X}_{tot}^k$ the probability density $P_\mathcal{S}^k$ in (3.10) is equal to one (its upper bound). In large sensor networks total coverage may be obtained by concentric configurations placed on and around $\partial\mathcal{S}$, as shown by the example in Figure B.1 with $k = 2$. The Lebesgue measure (or opening angle) of a finitely generated cone $K(\mathcal{D}_i, b^\ell) \in \mathbb{R}_+^2$, with origin b^ℓ, is bounded between 0 and π, and it is equal to π only when $b^\ell \in \mathcal{D}_i$. The Lebesgue measure on a union of cones, $\mu(\mathcal{K}_k(D, b^\ell))$, attains its upper bound π when all tracks through b^ℓ intersect k disks in D, and it is independent of k and N because all k-coverage cones are finitely generated. Then, the probability of detection for $\mathcal{X}_{tot}^k$ is obtained from (3.10) by letting $\mu(\mathcal{K}_k(D, b_y^\ell)) = \mu(\mathcal{K}_k(D, b_x^\ell)) = \mu(\mathcal{K}_k(D, b_{y'}^\ell)) = \mu(\mathcal{K}_k(D, b_{x'}^\ell)) = \pi\ \forall \ell$; i.e.,

$$P^k_{\mathcal{S}}(\mathcal{X}^k_{tot}) = \frac{\delta b}{4\pi L}\sum_{\ell=1}^{L/\delta b}(\pi+\pi) + \frac{\delta b}{4\pi L}\sum_{\ell=0}^{(L/\delta b-1)}(\pi+\pi)$$
$$= \frac{\delta b}{4\pi L}\left[\frac{L}{\delta b}2\pi + \frac{L}{\delta b}2\pi\right] = 1.$$

Appendix C. Performance analysis. Since every target remains in $\mathcal{S}$ and maintains $\mathcal{R}^j_\theta$ for the duration of the game and $V_{p_{\max}} > V_{\tau_{\max}}$, every target $j \in \mathcal{T}$ can be captured by a pursuer $j \in \mathcal{P}$ in time $t_{c^j_i}$, using the pursuit strategy in section 3.3, provided that $\mathcal{R}^j_\theta$ is fully observed. If $N \geq \lfloor 2L/r \rfloor$, an initial network configuration $\mathcal{X}^*_0$, defined in (3.2), with $\mathcal{P}^2_{\mathcal{S}}(\mathcal{X}^*_0) = 1$ can be obtained by solving a nonlinear program in which (3.10) is maximized with respect to $\mathcal{X}$, subject to $0 \leq x^i_p \leq L$ and $0 \leq y^i_p \leq L$ $\forall i \in I_{\mathcal{P}}$ [4]. As shown in Appendix B, in this total coverage configuration the network obtains at least two detections per target with probability one. After two detections are obtained, $\mathcal{R}^j_\theta$ is estimated and declared partially observed. Then, during every subsequent round, one sensor is moved from a cell κ_l to a cell $\kappa_\imath$ ($\kappa_l \to \kappa_\imath$), only if the overall probability of detection of the network, $(P_{\mathcal{R}}(\kappa_\imath) + \Delta P^k_{\mathcal{S}}(\kappa_l, \kappa_\imath))$, is increased by the change in configuration ($\mathcal{X}_l \to \mathcal{X}_\imath$). Since the probability density functions $P_{\mathcal{R}}$ and $P^k_{\mathcal{S}}$ are defined over the same set of targets $\mathcal{T}$ and the same search area $\mathcal{S}$, the probability of obtaining the $M(k-2)$ additional sensor detections required to declare all M target tracks fully observed increases at every round, provided that there are at least $(k-2)$ sensors in the network to obtain at least $(k-2)$ distinct detections per target. After a track $\mathcal{R}^j_\theta$ is fully observed, target j can always be captured by any sensor in the network, because $V_{p_{\max}} > V_{\tau_{\max}}$. Since, in the worst case, one sensor can be used to pursue every target sequentially, the game is guaranteed to terminate if $N \geq N_{\min}$, where

(C.1)
$$N_{\min} = \max\left\{\left\lfloor\frac{2L}{r}\right\rfloor + 1, k-2, 1\right\} = \frac{1}{2}\left[\left\lfloor\frac{2L}{r}\right\rfloor + 1 + k - 2 + \left|\left\lfloor\frac{2L}{r}\right\rfloor + 1 - k + 2\right|\right],$$

which simplifies to (4.1).

If $N < N_{\min}$, it can be shown by contradiction that the game cannot be guaranteed to terminate because there is a subset of tracks that may never be fully observed. From (C.1), there are two possible cases, namely, $N_{\min} = \lfloor 2L/r \rfloor + 1$ or $N_{\min} = k-2$, depending on the problem's parameters. In the first case, from the properties of the floor function, if $N < \lfloor 2L/r \rfloor + 1$, then $\lfloor N \rfloor < \lfloor 2L/r + 1 \rfloor$, and it follows that $N < 2L/r$, because N is an integer and $2L/r$ is a rational number (with the notable exception that if $2L/r$ is an integer, then $N_{\min}$ should be decreased by one). It also follows that $2Nr < 4L$, and thus for any $\mathcal{X}^*_0$ there exists at least one interval $\mathcal{I}_m = \{b \mid b \in \partial\mathcal{S},\ b \cap \mathcal{D}_i = \varnothing\ \forall i \in I_D\}$, as illustrated in Figure B.1, where if the dotted disk is removed, $N < 2L/r$. From Appendices A and B, the Lebesgue measure (or opening angle) μ of a coverage cone $K(\mathcal{D}_i, b)$ attains the upper bound π if and only if $b \in \mathcal{D}_i$. From (A.1), any ($k = 2$)-coverage cone $K_2(D^j_2, b)$ is the intersection of two coverage cones, e.g., $K(\mathcal{D}_i, b)$ and $K(\mathcal{D}_l, b)$, where $D^j_2 = \{\mathcal{D}_i, \mathcal{D}_l\} \subset D$, and thus $\mu(K_2(D^j_2, b)) \leq \min\{\mu(K(\mathcal{D}_i, b)), \mu(K(\mathcal{D}_l, b))\}$. It follows that for any intercept value $b^\ell \in \mathcal{I}_m$, $\mu(K(\mathcal{D}_i, b^\ell)) < \pi\ \forall i \in I_D$, and thus $\mu(K_2(D^j_2, b^\ell)) < \pi\ \forall j$, where K_2 is the ($k = 2$)-coverage cone for any 2-subset of D, as defined in (D.1). Thus, by Proposition 3.5, the set of tracks in the complement set $\mathcal{K}_m(b^\ell) = \mathcal{S} \setminus K_2(D^j_2, b^\ell)$, comprised of a union of cones, is detected by at most one sensor in $\mathcal{X}^*_0$ (Figure B.1).

Now, let $M = 1$ and assume that the target has a track $\mathcal{R}_\theta^m(b^\ell) \in \mathcal{K}_m(b^\ell)$, through an intercept $b^\ell \in \mathcal{I}_m$ (Figure B.1). Then, the target can be detected by at most one sensor in $\mathcal{X}_0^*$ and remains unobserved by definition. Since there are no other targets in $\mathcal{S}$, at every subsequent round in the game all sensors remain in detection mode, and all nodes in $\mathcal{G}$ remain void cells such that $P_\mathcal{R}(\kappa_\iota) = 0 \; \forall \kappa_\iota \in \mathcal{G}$. From (3.2), $\mathcal{X}_0^*$ is the configuration with the maximum value of $P_\mathcal{S}^2$. Thus, by its choice of κ_f for sensors in detection mode (section 3), the algorithm holds the sensors stationary at $\mathcal{X}_0^*$ at every round, and the game never terminates, because for this configuration $\mathcal{R}_\theta^m(b^\ell)$ always is unobserved. In the second case, $N_{\min} = k - 2 > \lfloor 2L/r \rfloor + 1$. Therefore, when $N < N_{\min}$, there may be a sufficient number of sensors in the network to obtain at least two detections per target and partially observe all tracks. However, after the tracks have been partially observed, there are not enough sensors to obtain the additional $k - 2$ independent detections required to declare any track fully observed. Thus, all sensors remain in detection mode, and the game never terminates because the targets are never captured. It can be concluded that if $N < N_{\min}$, the game cannot be guaranteed to terminate.

Based on the problem formulation in section 2, the time required to terminate the game T depends on the time required to obtain two detections per unobserved track, ΔT_d; the time required to obtain $1 < l < k$ additional detections per partially observed track, ΔT_o; and the time required to pursue every target after its track has been fully observed, ΔT_c. Although the method in section 3 allows different targets to be simultaneously detected and pursued, the worst-case scenario is one where there is no overlap between these time periods, i.e., $T = \Delta T_d + \Delta T_o + \Delta T_c$.

After all targets are in $\mathcal{S}$, the maximum time required by a network with $N_{\min}$ sensors, positioned at $\mathcal{X}_0^*$, to obtain at least two detections per target is the time required by the slowest target to travel the longest distance in $\mathcal{S}$ between two disks $\mathcal{D}_i$ and $\mathcal{D}_j$ with p_i and p_j at opposite corners, i.e., $\Delta T_d(N_{\min}) = (\sqrt{2}L - 2r)/V_{\tau_{\min}}$. Consider now a network with $N \geq N_{\min}$ sensors. Before obtaining any detections, it can be easily shown that the configuration that maximizes the distance between any two sensors is one where they are packed concentrically in $\mathcal{S}$ (e.g., the grey disks in Figure B.1). By the definition of a triangular number, this configuration can be achieved by increasing the number of sensors according to

$$N = \ell \left\lfloor \frac{2L}{r} \right\rfloor - 8\mathrm{T}_{(\ell-1)} = \ell \left\lfloor \frac{2L}{r} \right\rfloor - 4\ell(\ell - 1), \quad \ell = 1, \dots, \left\lfloor \frac{L}{4r} \right\rfloor, \tag{C.2}$$

where ℓ is an integer and $\lfloor L/4r \rfloor$ is the maximum value of ℓ that allows one to pack N sensors in $\mathcal{S}$. T_n denotes the triangular number of n and is used to represent the decrease in the number of sensors that can be placed at the corners as N increases (see Figure B.1). Then, the maximum time required to obtain two distinct detections is the time required by the slowest target to travel the maximum distance between any two sensors, that is,

$$\begin{aligned} \Delta T_d(N) &= \frac{1}{V_{\tau_{\min}}} \max \left\{ 2r, \sqrt{2}L - 2r[2\sqrt{2}(\ell - 1) + 1] \right\} \\ &= \frac{1}{V_{\tau_{\min}}} \frac{1}{2} \left\{ 2r + \sqrt{2}L - 2r[2\sqrt{2}(\ell - 1) + 1] - \left| 2r - \sqrt{2}L + 2r[2\sqrt{2}(\ell - 1) + 1] \right| \right\} \\ &= \frac{1}{V_{\tau_{\min}}} \left\{ \frac{\sqrt{2}}{2}L - 2\sqrt{2}r(\ell - 1) + \left| 2r[1 + \sqrt{2}(\ell - 1)] - \frac{\sqrt{2}}{2}L \right| \right\}. \end{aligned} \tag{C.3}$$

After all targets have been partially observed, $(k-2)$ additional detections per target are required to declare them fully observed (where it is assumed that $k > 2$). If there are $N_{\min}$ sensors in the network, this can always be accomplished by repeatedly moving the sensors until all M target tracks are fully observed. Since the maximum distance that must be traveled by any of the sensors to obtain an additional detection is $(\sqrt{2}L - r)$, the time required to fully observe all M target tracks is

$$\Delta T_o(N_{\min}) = \left[\left\lfloor \frac{(k-2)M}{N_{\min}} \right\rfloor + 1 \right] \frac{(\sqrt{2}L - r)}{\bar{V}_p}. \tag{C.4}$$

However, if at least $(k-2)M$ sensors are available, they can all be moved at once to each obtain one additional detection, and, assuming that the running time to reconfigure them is negligibly small, all M target tracks are fully observed after a time $(\sqrt{2}L - r)/\bar{V}_p$.

Suppose that there are $\lfloor 2L/r \rfloor \leq N \geq M$ sensors available to pursue the targets after they have been fully observed. Then, the maximum distance that must be traveled to capture target j by switching the nearest sensor to pursuit mode is the distance between the sensor that has obtained the last kth detection from j and the interception point δ. For the purpose of analysis, one can place an inertial frame of reference at $\tau_j(t_k)$ and align it with the target track such that $x^j_\tau(t_k) = y^j_\tau(t_k) = \theta^j_\tau = 0$. Then, at time t_k the sensor must be within a distance r from the target position, with a heading θ_p. From (3.16), assuming that the turn time is negligibly small, the capture time t_c must satisfy the equality

$$\begin{aligned} (V_{p_{\max}} t_c)^2 &= r^2 \cos^2 \theta_p + (\bar{V}_\tau t_c)^2 - 2 t_c r \bar{V}_\tau \cos \theta_p + r^2 \sin^2 \theta_p \\ &= (\bar{V}_\tau t_c)^2 - 2 t_c r \bar{V}_\tau \cos \theta_p + r^2 \end{aligned} \tag{C.5}$$

for any θ_p, since the sensors are assumed to travel at their maximum speed when in pursuit mode (section 3.3). Differentiating (C.5) with respect to θ_p and setting the result equal to zero, it can be shown that the heading with maximum capture time is $\theta^*_p = \pi$ with respect to the target track. Thus, the maximum capture time can be obtained from

$$t^*_c = \frac{1}{V_{p_{\max}}} \left\| \begin{matrix} (r \cos \theta^*_p - t^*_c \bar{V}_\tau) \\ r \sin \theta^*_p \end{matrix} \right\| \tag{C.6}$$

by letting $\theta^*_p = \pi$ and solving for $t^*_c = r/(V_{p_{\max}} - \bar{V}_\tau)$. Since all of the M sensors that obtained the last kth detection from the M targets can be deployed simultaneously to pursue them, the maximum time required to capture them is $\Delta T_c(N) = r/(V_{p_{\max}} - \bar{V}_\tau)$.

If there are only $N_{\min}$ sensors in the network, they need to be deployed $(\lfloor M/N_{\min} \rfloor + 1)$ times in order to capture all M targets that have been fully observed. The first time, the $N_{\min}$ sensors that obtained the last kth detection from $N_{\min}$ targets are deployed in $r/(V_{p_{\max}} - \bar{V}_\tau)$ maximum time, as shown in the previous paragraph. However, when the sensors are subsequently redeployed, they could be anywhere in $\mathcal{S}$, since they reenter the game after having captured other targets. It can be shown by solving a simple constrained-optimization problem (omitted here for brevity) that for a sensor and a target located anywhere in $\mathcal{S}$ the maximum capture time is $L[\bar{V}_\tau + (2V^2_{p_{\max}} - \bar{V}^2_\tau)^{1/2}]/(V^2_{p_{\max}} - \bar{V}^2_\tau)$. Thus, it follows that the maximum capture time with $N_{\min}$ sensors is

$$\Delta T_c(N_{\min}) = \frac{r}{(V^2_{p_{\max}} - \bar{V}^2_\tau)} + \left\lfloor \frac{M}{N_{\min}} \right\rfloor \frac{\left(\bar{V}_\tau + \sqrt{2V^2_{p_{\max}} - \bar{V}^2_\tau}\right)}{(V^2_{p_{\max}} - \bar{V}^2_\tau)^2} L. \tag{C.7}$$

From the above analysis, the minimum number of sensors $N_{\min}$, obtained by the maximum of all minima in (4.1), is required to terminate the game in a maximum time T_u in (4.2). Also, the maximum time required to end the game can be reduced to T_ℓ, in (4.5), by increasing the number of sensors to the maximum of all maxima,

$$\text{(C.8)} \qquad N_\ell = \max\left\{\ell\left\lfloor\frac{2L}{r}\right\rfloor - 4\ell(\ell-1), (k-2)M, M\right\}, \quad \ell = 1, \dots, \left\lfloor\frac{L}{4r}\right\rfloor,$$

which simplifies to (4.4), since $k > 2$.

Appendix D. Complexity analysis. During every detection round, the main computations required by the methodology in section 3 are the cell-decomposition procedure to obtain $\mathcal{G}$, the computation of the sensing reward (3.1) for every observation cell in $\mathcal{G}$, and the search for the optimal channel (3.3), by the A* algorithm. Therefore, we first analyze the complexity of the cell-decomposition procedure. Since the sensor field-of-view $\mathcal{D}_i$ is a disk, the decomposition may involve generalized polygons. To avoid this case, $\mathcal{D}_i$ can be approximated by a regular octagon $\hat{\mathcal{D}}_i$ tightly contained in $\mathcal{D}_i$. For simplicity, θ_p^i is assumed fixed. The obstacle-free configuration space $\mathcal{C}_{free}^i$ for the ith pursuer $\mathcal{A}_i$ can be decomposed via the method in section 3.2. The position $p_i = (x_p^i,\ y_p^i)$ of $\mathcal{A}_i$ is abbreviated here by $(x,\ y)$, and $\mathcal{C}$ denotes the robot configuration space, which can be assumed to be a rectangle or a union of rectangles. Let $\kappa = [x_\kappa, x'_\kappa] \times [y_\kappa, y'_\kappa]$ denote a rectangle in $\mathbb{R}^2$. From the problem formulation, $\mathcal{A}_i$, $\hat{\mathcal{D}}_i$, and $\mathcal{O}_k$, $k = 1, \dots, n$, are assumed to be convex polygons, and targets $\mathcal{R}_j$, $j = 1, \dots, l$, are rays in the workspace $\mathcal{S}$; then it follows that $\mathcal{CO}_k$, $k = 1, \dots, n$, and thus $\mathcal{CR}_j$, $j = 1, \dots, l$, all are convex [29]. Let $n_{e_\mathcal{A}}$ denote the number of edges describing $\mathcal{A}_i$, which is assumed to be a fixed constant. The number of edges defining sensor $\hat{\mathcal{D}}_i$ is eight. The running time to compute the two-dimensional C-obstacle $\mathcal{CO}_k$, for all $k = 1, \dots, n$, in $\mathcal{S}$ is $O(n_{e_\mathcal{O}})$. The running time to compute $\mathcal{CR}_j$, for all $j = 1, \dots, l$, in $\mathcal{S}$ is $O(n_{e_\mathcal{R}})$ [29, 31]. The decomposition of the complement of $\mathcal{CO}_k$, $k = 1, \dots, n$, and $\mathcal{CR}_j$, $j = 1, \dots, l$, into convex cells is a vertical decomposition in two dimensions [29, 31]. Thus, the complexity of the decomposition of $\mathcal{C}_{void}^i$, in step (I), is $O((n_{e_\mathcal{O}} + n_{e_\mathcal{R}})\log(n_{e_\mathcal{O}} + n_{e_\mathcal{R}}))$. In step (II), the decomposition of $\mathcal{CR}_j \setminus \bigcup_{j=1}^n \mathcal{CO}_k$, for each $j = 1, \dots, p$, can be also implemented via vertical decomposition in two dimensions [29, 31] with complexity $O(n_{e_\mathcal{O}} \log n_{e_\mathcal{O}})$. Thus, the complexity of step (II) is $O(n_{e_\mathcal{R}} n_{e_\mathcal{O}} \log n_{e_\mathcal{O}})$. Therefore, the overall complexity of the cell-decomposition method is $O((n_{e_\mathcal{O}} + n_{e_\mathcal{R}})\log(n_{e_\mathcal{O}} + n_{e_\mathcal{R}})) + O(n_{e_\mathcal{R}} n_{e_\mathcal{O}} \log n_{e_\mathcal{O}}) = O((n_{e_\mathcal{O}} + n_{e_\mathcal{R}})\log(n_{e_\mathcal{O}} + n_{e_\mathcal{R}}) + n_{e_\mathcal{R}} n_{e_\mathcal{O}} \log n_{e_\mathcal{O}})$, whereas the complexity of a decomposition involving only obstacles is $O(n_{e_\mathcal{O}} \log(n_{e_\mathcal{O}}))$ [29, 31]. By using the *approximate-and-decompose* method in [52], the cell decomposition in section 3.2 can be run in $O((n_{e_\mathcal{O}} + n_{e_\mathcal{R}})^2)$ [8].

Next, we analyze the complexity of the probability of cooperative detections (3.10), which is clearly the most expensive computation in the sensing reward (3.1). Let $b_y = \ell \cdot \delta b$, where $\ell = 0, \dots, L/\delta b$. Consider the complexity of computing (B.5) and denote it by Γ_y^k. Then, the time complexity of computing (3.10) is $O(4\Gamma_y^k)$. All tracks $\mathcal{R}_\alpha(b_y)$ detected by a set of k sensors, D_k, are contained by the k-coverage cone of D_k, and those detected by at least k sensors in a set of n sensors, D, are given by the union in (B.1). By computing the geometric union of the convex cones $K_k(D_k^j, b_y), j = 1, \dots, m$, K_k can equivalently be expressed as a union of $\wp$ disjoint convex cones,

$$\mathcal{K}_k(D, b_y) = \bigcup_{\jmath=1}^{\wp} cone(\hat{l}_\jmath^*, \hat{h}_\jmath^*), \tag{D.1}$$

where $cone(\hat{l}_\jmath^*, \hat{h}_\jmath^*) \bigcap cone(\hat{l}_\imath^*, \hat{h}_\imath^*) = \varnothing \, \forall \jmath \neq \imath$, where $\imath, \jmath = 1, \dots, \wp$, and $\hat{l}_\jmath^* \prec \hat{l}_\jmath^* \prec \hat{l}_\imath^* \, \forall \jmath < \imath$. Clearly, $\wp \leq m$, and $\hat{l}_\jmath^* \in \{\hat{l}_j^* | j = 1, \dots, m\}$ and $\hat{h}_\jmath^* \in \{\hat{h}_j^* | j = 1, \dots, m\}, \jmath = 1, \dots, \wp$. Therefore, (B.5) can be obtained by two steps: (i) obtain $K_k(\mathcal{D}, b_y)$ in (D.1), and (ii) obtain (B.5) from the following equations:

$$\Pr\{D_{jk} = 1, \mathcal{R}_\theta^j(b_y^\ell \in \partial\mathcal{S})\} = \sum_{\ell=0}^{(L/\delta b - 1)} \sum_{\jmath=1}^{\wp} \sin^{-1} ||\hat{l}_\jmath^* \times \hat{h}_\jmath^*||. \tag{D.2}$$

The computation of the inner summation above can be performed in two steps. First, for every D_k^j, $j = 1, \dots, m$, $\forall i \in I_{D_k^j}$, compute $\sin\alpha_i$ in (3.6) and $\cos\alpha_i = (1 - \sin{\alpha_i}^2)^{1/2}$, then compute $\hat{h}_i$ in (3.4) and $\hat{l}_i$ in (3.5); choose optimal $\hat{l}_j^*$ and $\hat{h}_j^*$ so that $\hat{l}_j^* = \sup\{\hat{l}_i | i \in I_{D_k^j}\}$ and $\hat{h}_j^* = \inf\{\hat{h}_j | \, j \in I_{D_k^j}\}$. Then, $\mathcal{K}_k(D, b_y)$ is obtained in the form of (D.1); then compute the measure in (D.2).

Assume that the running time to compute the elementary function $\sin\alpha_i$ in (3.6) is a constant. The complexity of its inverse $\sin^{-1}\alpha_i$ is also a constant, since all elementary functions are analytic and hence invertible by means of Newton's method. The time to compute all $\sin\alpha_i$ and $\cos\alpha_i$, $i \in I_{D_k^j}$, is $O(k) + O(k)$, i.e., $O(k)$. Then, $\hat{h}_i$ and $\hat{l}_i$ can be obtained by simple multiplication and addition operations. Since there is no need to order $\{\hat{l}_i | i \in I_{\mathcal{D}_k^j}\}$ and $\{\hat{h}_j | \, j \in I_{\mathcal{D}_k^j}\}$, $\hat{l}_j^*$ and $\hat{h}_j^*$ can be obtained in linear running time $O(k)$. The complexity to generate m convex cones $cone(\hat{l}_j^*, \hat{h}_j^*) \; \forall j = 1, \dots, m$ is $O(mk)$. The computation of the unions of finite convex cones is exactly similar to the computation of the union of finite closed intervals in $\mathbb{R}^1$. This class of problems is well known as Klee's measure problem [26]. In 1977, Klee considered the following problem: given a collection of m intervals in the real line, compute the length of their union; he then presented an algorithm [26] to solve this problem with computational complexity (or "running time") $O(m \log m)$. Fredman and Weide showed that Klee's algorithm, based on sorting the intervals, was optimal [20]. Therefore, the complexity of the second step is $O(m \log m)$, and $\Gamma_y^k = n_\delta(O(mk) + O(m \log m)) = O(n_\delta m(k + \log m))$, where $n_\delta = L/\delta b$ and m is the binomial coefficient N choose k. Finally, the time complexity of (3.10) is $4 \cdot \Gamma_y^k$ for every observation cell, and thus the required to compute (3.1) for all $n_{\underline{\mathcal{K}}}$ observation cells in $\mathcal{G}$ is $O(n_{\underline{\mathcal{K}}} n_\delta m(k + \log m))$.

Finally, $\mathcal{G}$ is searched for the optimal channel (3.3) using the A* algorithm. The time complexity of A* depends on the heuristic function and, therefore, on the cost or reward function attached to the arcs of $\mathcal{G}$. For the reward (3.1), the easiest choice of heuristic function is $h(x) = 0 \; \forall x$. Then, the A* reduces to Dijkstra's algorithm [16], for which the running time is $O(n_\mathcal{G}^2 + n_{\mathcal{G}_a})$. Therefore, the time complexity of a detection round is given by the leading term in (4.6), which depends on the characteristics of the workspace, robotic sensors, and targets. When the game round consists of deploying a sensor in pursuit mode, the optimal channel μ_p^* is computed by first determining the interception point and time, δ and $t_{c_i^j}$, and then using the A* algorithm to find the shortest path from $p_i(t_0)$ to δ in $\mathcal{G}$. δ and $t_{c_i^j}$ are computed by solving four nonlinear equations in four variables, using Newton's method [34].

Since the gradient evaluation equals 4^2 component function evaluations, the time complexity of one Newton's method iteration is $O(4^3)$. Although the number of iterations required is unknown, Newton's method is known to converge locally in linear, or even superlinear, time. Therefore, the time complexity for computing the pursuit strategy is that of the A*. Since, in this case, the cost attached to the arcs of $\mathcal{G}$ is the Euclidean distance (3.17), the heuristic function $h(x)$ can be chosen as the straight-line distance, and the complexity can be reduced to $O((n_{\mathcal{G}_a} + n_{\mathcal{G}}) \log_2 n_{\mathcal{G}})$ [44].

REFERENCES

[1] E. U. Acar, *Path planning for robotic demining: Robust sensor-based coverage of unstructured environments and probabilistic methods*, Internat. J. Robotic Res., 22 (2003), pp. 7–8.

[2] M. Adler, H. Räcke, N. Sivadasan, C. Sohler, and B. Vöcking, *Randomized pursuit-evasion in graphs*, Combin. Probab. Comput., 12 (2003), pp. 225–244.

[3] A. Arsie and E. Frazzoli, *Efficient routing of multiple vehicles with no explicit communications*, Internat. J. Robust Nonlinear Control, 18 (2007), pp. 154–164.

[4] K. C. Baumgartner and S. Ferrari, *A geometric transversal approach to analyzing track coverage in sensor networks*, IEEE Trans. Comput., 57 (2008), pp. 1113–1128.

[5] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*, Wiley Interscience, Hoboken, NJ, 2006.

[6] D. P. Bertsekas, *Convex Analysis and Optimization*, Athena Scientific, Belmont, MA, 2003.

[7] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 2007.

[8] C. Cai and S. Ferrari, *Information-driven sensor path planning by approximate cell decomposition*, IEEE Trans. Syst. Man Cyber. Part B, 39 (2009), pp. 1–18.

[9] H. Choset, *Coverage for robotics: A survey of recent results*, Ann. Math. Artif. Intell., 31 (2001), pp. 113–126.

[10] J. Clark and R. Fierro, *Mobile robotic sensors for perimeter detection and tracking*, ISA Trans., 46 (2007), pp. 3–13.

[11] J. Cortés, S. Martínez, T. Karatas, and F. Bullo, *Coverage control for mobile sensing networks*, IEEE Trans. Robotics Automat., 20 (2004), pp. 243–255.

[12] I. J. Cox and M. L. Miller, *On finding ranked assignments with application to multitarget tracking and motion correspondence*, IEEE Trans. Aerospace Electronic Systems, 31 (1995), pp. 486–489.

[13] D. Cruz, J. McClintock, B. Perteet, O. Orqueda, Y. Cao, and R. Fierro, *Decentralized cooperative control: A multivehicle platform for research in networked embedded systems*, IEEE Control Syst. Mag., 27 (2007), pp. 58–78.

[14] R. V. Dam, *Soil effects on thermal signatures of buried nonmetallic landmines*, in Detection and Remediation Technologies for Mines and Minelike Targets VIII, Proc. SPIE 5089, SPIE, Bellingham, WA, 2003, pp. 1210–1218.

[15] H. F. Davis and A. D. Snider, *Vector Analysis*, Wm. C. Brown, Dubuque, IA 1987.

[16] E. W. Dijkstra, *A note on two problems in connexion with graphs*, Numer. Math., 1 (1959), pp. 269–271.

[17] S. Ferrari, *Track coverage in sensor networks*, in Proceedings of the IEEE American Control Conference, Minneapolis, MN, 2006, IEEE Press, Piscataway, NJ, pp. 2053–2059.

[18] S. Ferrari and A. Vaghi, *Demining sensor modeling and feature-level fusion by Bayesian networks*, IEEE Sensors, 6 (2006), pp. 471–483.

[19] R. Fierro, A. Das, J. Spletzer, J. Esposito, V. Kumar, J. P. Ostrowski, G. Pappas, C. J. Taylor, Y. Hur, R. Alur, I. Lee, G. Grudic, and J. Southall, *A framework and architecture for multi-robot coordination*, Internat. J. Robotic Res., 21 (2002), pp. 977–995.

[20] M. L. Fredman and B. Weide, *On the complexity of computing the measure of* $\cup[a_i, b_i]$, Comm. ACM, 21 (1978), pp. 540–544.

[21] A. Gad, F. Majdi, and M. Farooq, *A comparison of data association techniques for target tracking in clutter*, in Proceedings of the Fifth IEEE International Conference on Information Fusion, 2002, IEEE Press, Piscataway, NJ, vol. 2, pp. 1126–1133.

[22] A. Ganguli, J. Cortés, and F. Bullo, *Maximizing visibility in nonconvex polygons: Nonsmooth analysis and gradient algorithm design*, SIAM J. Control Optim., 45 (2006), pp. 1657–1679.

[23] J. E. Goodman, R. Pollack, and R. Wenger, *Geometric transversal theory*, in New Trends in Discrete and Computational Geometry, J. Pach, ed., Springer-Verlag, New York, Berlin, 1991, pp. 163–198.

[24] V. Isler, S. Kannan, and S. Khanna, *Randomized pursuit-evasion in a polygonal environment*, IEEE Trans. Robotics, 21 (2005), pp. 875–884.

[25] P. Juang, H. Oki, Y. Wang, M. Martonosi, L. Peh, and D. Rubenstein, *Energy efficient computing for wildlife tracking: Design tradeoffs and early experiences with zebranet*, in Proceedings of the 10th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS-X), 2002, ACM, New York, pp. 96–108.

[26] V. Klee, *Can the measure of $\cup[a_i, b_i]$ be computed in less than $O(n \log n)$ steps?*, Amer. Math. Monthly, 84 (1977), pp. 284–285.

[27] S. Koenig, C. Tovey, and Y. Smirnov, *Performance bounds for planning in unknown terrain*, Artificial Intelligence, 147 (2003), pp. 253–279.

[28] A. S. LaPaugh, *Recontamination does not help search a graph*, J. ACM, 40 (1993), pp. 224–245.

[29] J. C. Latombe, *Robot Motion Planning*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1991.

[30] J. C. Latombe, A. Lazanas, and S. Shekhar, *Robot motion planning with uncertainty in control and sensing*, Artificial Intelligence, 52 (1991), pp. 1–47.

[31] S. M. LaValle, *Planning Algorithms*, Cambridge University Press, London, 2006.

[32] A. Lazanas and J. C. Latombe, *Motion planning with uncertainty—A landmark approach*, Artificial Intelligence, 76 (1995), pp. 287–317.

[33] J. Leonard, H. Durrant-Whyte, and I. Cox, *Dynamic map building for an autonomous mobile robot*, Internat. J. Robotic Res., 11 (1992), pp. 286–298.

[34] J. J. Moré and M. Y. Cosnard, *Numerical solution of nonlinear equations*, ACM Trans. Math. Software, 5 (1999), pp. 64–85.

[35] G. Oriolo, G. Ulivi, and M. Vendittelli, *Real-time map building and navigation for autonomous robots in unknown environments*, IEEE Trans. Syst. Man Cyber., 28 (1995), pp. 316–333.

[36] J. O'Rourke, *Art Gallery Theorems and Algorithms*, Oxford University Press, London, 1987.

[37] J. Paik, *Image processing-based mine detection techniques using multiple sensors: A review*, Subsurface Sensing Technol. Appl., 3 (2002), pp. 203–252.

[38] T. D. Parsons, *Pursuit-evasion in a graph*, in Theory and Applications of Graphs, Y. Alavi and D. R. Lick, eds., Lecture Notes in Math. 642, Springer-Verlag, Berlin, 1978, pp. 426–441.

[39] A. B. Poore and N. Rijavec, *A numerical study of some data association problems arising in multitarget tracking*, Comput. Optim. Appl., 3 (1994), pp. 27–57.

[40] M. Qian and S. Ferrari, *Probabilistic deployment for multiple sensor systems*, in Proceedings of the 12th SPIE Symposium on Smart Structures and Materials: Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems, vol. 5765, San Diego, 2005, pp. 85–96.

[41] N. Rao, *Robot navigation in unknown generalized polygonal terrains using vision sensors*, IEEE Trans. Syst. Man Cyber., 25 (1995), pp. 947–962.

[42] N. Rao, S. Hareti, W. Shi, and S. Iyengar, *Robot Navigation in Unknown Terrains: Introductory Survey of Non-heuristic Algorithms*, Technical Report ORNL/TM-12410, Oak Ridge National Laboratory, Oak Ridge, TN, 1993.

[43] S. M. Ross, *Introduction to Stochastic Dynamic Programming*, Academic Press, Orlando, FL, 1983.

[44] S. Russell and P. Norvig, *Artificial Intelligence, A Modern Approach*, Prentice–Hall, Upper Saddle River, NJ, 2003.

[45] T. Shermer, *Recent results in art galleries*, Proc. IEEE, 80 (1992), pp. 1384–1399.

[46] S. Thurn, *Learning metric-topological maps for indoor mobile robot navigation*, Artificial Intelligence, 99 (1998), pp. 21–71.

[47] J. Urritia, *Art gallery and illumination problems*, in Handbook on Computational Geometry, J. Sack and J. Urritia, eds., Elsevier Science, New York, 1992, pp. 387–434.

[48] R. Vidal, O. Shakernia, J. Kim, D. Shim, and S. Sastry, *Probabilistic pursuit-evasion games: Theory, implementation, and experimental evaluation*, IEEE Trans. Robotics Automat., 18 (2002), pp. 662–669.

[49] T. A. Wettergren, *Performance of search via track-before-detect for distributed sensor networks*, IEEE Trans. Aerospace Electronic Systems, 44 (2008), pp. 314–325.

[50] T. A. Wettergren, R. L. Streit, and J. R. Short, *Tracking with distributed sets of proximity sensors using geometric invariants*, IEEE Trans. Aerospace Electronic Systems, 40 (2004), pp. 1366–1374.

[51] P. Yang, R. Freeman, and K. Lynch, *Distributed cooperative active sensing using consensus filters*, in Proceedings of the IEEE International Conference on Robotics and Automation, Rome, 2007, IEEE Press, Piscataway, NJ, pp. 405–410.

[52] D. Zhu and J.-C. Latombe, *New heuristic algorithms for efficient hierarchical path planning*, IEEE Trans. Robotics Automat., 7 (1991), pp. 9–20.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 1, pp. 321–350

© 2009 Society for Industrial and Applied Mathematics

SEQUENTIAL LOCALIZATION OF SENSOR NETWORKS[*]

J. FANG[†], M. CAO[‡], A. S. MORSE[†], AND B. D. O. ANDERSON[§]

Abstract. The sensor network localization problem with distance information is to determine the positions of all sensors in a network, given the positions of some of the sensors and the distances between some pairs of sensors. A definition is given for a sensor network in the plane to be "sequentially localizable." It is shown that the graph of a sequentially localizable network must have a "bilateration ordering," and a polynomial time algorithm is given for deciding whether or not a network's graph has such an ordering. A provably correct algorithm is given which consists of solving a sequence of quadratic equations, and it is shown that the algorithm can localize any localizable network in the plane whose graph has a bilateration ordering.

Key words. sensor networks, localization, graph theory, rigidity

AMS subject classifications. 68W01, 68R99

DOI. 10.1137/070679144

1. Introduction.

Determining the positions of sensors is essential in many network applications such as geographic routing, coverage and creating formations. Equipping each sensor in a network with GPS is not feasible in many cases because of the large number of sensors and the cost associated with a GPS unit. Hence, we attack this problem by exploiting the connectivity of a sensor network and some common capabilities of sensors. More specifically, we assume a sensor can measure its distances to and communicate with certain other sensors in the network. The sensor network localization problem with distance information is to determine the positions of all sensors in a network given the positions of some sensors and the distances between some pairs of sensors. A sensor whose position is given is called an *anchor*. A network in $\mathbb{R}^d$ is said to be *localizable* if there exists exactly one position in $\mathbb{R}^d$ corresponding to each nonanchor sensor such that the given intersensor distances are satisfied. The authors of [5] use rigidity theory to give the necessary and sufficient conditions for a network to be localizable. However, the process of localizing a network has been shown to be NP-hard even when the network is known to be localizable [2]. This leaves us with the more refined questions of how we should go about localizing networks, and what kinds of networks can we "efficiently" localize.

The characterization of networks which can be "easily" or "efficiently" localized is not complete, even for the ideal case where the given distance measurements are exact. In [14], global nonlinear optimization techniques combined with heuristics are

[*] Received by the editors January 3, 2007; accepted for publication (in revised form) September 2, 2008; published electronically February 11, 2009. The work of Fang, Cao, and Morse was supported in part by grants from the U.S. Army Research Office and the U.S. National Science Foundation and by a gift from the Xerox Corporation. The work of Anderson was supported by the Research School of Information Sciences and Engineering and National ICT Australia which is funded by the Australian Government's Department of Communications, Information Technology and the Arts and the Australian Research Council through the Backing Australia's Ability initiative and the ICT Center of Excellence Program.

http://www.siam.org/journals/sicon/48-1/67914.html

[†] Department of Electrical Engineering, Yale University, New Haven, CT 06520 (jia.fang@yale.edu, as.morse@yale.edu).

[‡] Faculty of Mathematics and Natural Sciences, ITM, University of Groningen, Groningen, The Netherlands (m.cao@rug.nl).

[§] Australian National University and National ICT Australia Ltd., 216 Northbourne Ave., Canberra ACT 2601 Australia (Brian.Anderson@nicta.com.au).

used to estimate sensor information. In [13], a "fold-free" layout of the network is first estimated, and then force-based relaxation methods are used to refine the estimated sensor positions. In [16, 15], the distance between each pair of sensors is estimated from the given intersensor distances using a shortest path algorithm, and classical multidimensional scaling techniques are used to assign positions to each sensor to approximate the given distance information. In [3], a semidefinite programming–based algorithm is given for a class of dense networks when a sufficiently large number of intersensor distances are known. In this work, we are interested in provably correct localization algorithms and the kinds of networks that can be "efficiently" localized by them. We assume the given intersensor distances are exact distance measurements.

The characterization of efficiently localizable networks has been investigated in [1] and we extend the results of that paper. We present a localization algorithm called "Sweeps" which consists of solving a sequence of a finite number of quadratic equations, where the solution of each equation is easily obtainable by the well-known quadratic formula. We give a simple graphical characterization of all networks which can be localized by Sweeps, and we use graph rigidity theory to give some graphical characterizations of networks that can be efficiently localized by Sweeps. We also introduce the concept of "sequential" localization algorithms, and we say a network is *sequentially localizable* if it can be localized by some sequential localization algorithm. We show that Sweeps is a sequential localization algorithm which can localize *all* sequentially localizable networks. The Sweeps algorithm we present in this work is limited because we assume that the given intersensor distance measurements are exact. We refer the interested reader to [6] in which we adapt the Sweeps algorithm for the case of inaccurate distance measurements.

In section 2, we review the theoretical background of the localization problem from graph rigidity theory, and we give the terms and definitions to be used in the exposition that follows. In section 3, we introduce the notions of "bilateration orderings" and "sequentially localizable" networks. In section 4 we present the Restricted Sweeps algorithm on which the Sweeps algorithm is based, and in section 5 we present the Sweeps algorithm. In section 6 we characterize the class of networks localizable by Sweeps, and we show that all sequentially localizable networks are localizable by Sweeps. In section 7, we characterize some classes of networks which can be "efficiently" localized by either Sweeps or Restricted Sweeps, and in section 8 we characterize some classes of networks which are localizable by Restricted Sweeps. We conclude with future work and research problems in section 9.

2. Background.

A *multipoint* $p = \{p_1, \dots, p_n\}$ in d-dimensional space is a set of n points in $\mathbb{R}^d$ labeled $p_1, \dots, p_n$. Because we are concerned only with networks in the plane, we will henceforth restrict our attention to the case of $d = 2$. Two multipoints $p = \{p_1, \dots, p_n\}$ and $q = \{q_1, \dots, q_n\}$ of n points in $\mathbb{R}^2$ are *congruent* if for all $i, j \in \{1, \dots, n\}$, the distance between p_i and p_j is equal to the distance between q_i and q_j. A graph with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$ is denoted $(\mathcal{V}, \mathcal{E})$. A *simple graph* is a graph for which there is at most one edge between any two distinct vertices, and no edge between a vertex and itself. A *point formation* of n points at a multipoint $p = \{p_1, \dots, p_n\}$ consists of p and a simple undirected graph $\mathbb{G}$ with vertex set $\mathcal{V} = \{1, \dots, n\}$, and is denoted by $(\mathbb{G}, p)$. If (i, j) is an edge in $\mathbb{G}$, then the *length of edge* (i, j) in the point formation $(\mathbb{G}, p)$ is the distance between p_i and p_j. Two point formations with the same graph have the same edge lengths in the case when the length of each edge in the graph is the same in both point formations.

For $t \in \mathbb{R}^2$, let $\|t\|$ denote the Euclidean norm on $\mathbb{R}^2$. For any multipoint $p =$

$\{p_1, \ldots, p_n\}$ in $\mathbb{R}^2$ and $\epsilon > 0$, let $\mathcal{B}_p(\epsilon)$ denote the set of all multipoints $q = \{q_1, \ldots, q_n\}$ in $\mathbb{R}^2$, where $\|p_i - q_i\| < \epsilon$ for all $i \in \{1, \ldots, n\}$. A point formation $(\mathbb{G}, p)$ is *rigid* in $\mathbb{R}^2$ if there exists $\epsilon > 0$ such that for all $q \in \mathcal{B}_p(\epsilon)$, p and q are congruent whenever $(\mathbb{G}, p)$ and $(\mathbb{G}, q)$ have the same edge lengths. Roughly speaking, a rigid point formation is one that cannot be continuously deformed without causing an edge length to change. A graph $\mathbb{G}$ is said to be *rigid* in $\mathbb{R}^2$ if there exists a multipoint p in $\mathbb{R}^2$ and $\epsilon > 0$ such that $(\mathbb{G}, q)$ is rigid in $\mathbb{R}^2$ for all $q \in \mathcal{B}_p(\epsilon)$. A set consisting of a finite number of elements from $\mathbb{R}$ is said to be *algebraically independent over the rationals* if its elements do not satisfy any nonzero multivariable polynomial equation with rational coefficients. A multipoint is said to be *generic* if the set consisting of the coordinates of its points is algebraically independent over the rationals. It is known that if a multipoint p is generic, then a point formation $(\mathbb{G}, p)$ is rigid if and only if $\mathbb{G}$ is rigid.

A point formation $(\mathbb{G}, p)$ in $\mathbb{R}^2$ is *globally rigid* in $\mathbb{R}^2$ if multipoints p and q are congruent whenever $(\mathbb{G}, p)$ and $(\mathbb{G}, q)$ have the same edge lengths. In other words, edge lengths of a globally rigid point formation uniquely determine all intervertex distances. A graph $\mathbb{G}$ is said to be *globally rigid* in $\mathbb{R}^2$ if there exist multipoint p in $\mathbb{R}^2$ and $\epsilon > 0$ such that $(\mathbb{G}, q)$ is globally rigid in $\mathbb{R}^2$ for all $q \in \mathcal{B}_p(\epsilon)$. It is known that if a multipoint p in $\mathbb{R}^2$ is generic, then the point formation $(\mathbb{G}, p)$ is globally rigid in $\mathbb{R}^2$ if and only if $\mathbb{G}$ is globally rigid in $\mathbb{R}^2$. For any integer $k > 1$, a graph is said to be *k connected* if there does not exist a set of $k - 1$ vertices whose removal disconnects the graph. It is known that a graph with four or more vertices is globally rigid in $\mathbb{R}^2$ if and only if the graph is three connected and there does not exist an edge of the graph whose removal results in a graph which is not rigid in $\mathbb{R}^2$ [10, 9, 4]. There are a number of polynomial time algorithms such as Pebble Game for determining if a graph is rigid in $\mathbb{R}^2$ [11]. Since the k connectedness of a graph can also be efficiently determined, it follows that the global rigidity of a graph in $\mathbb{R}^2$ can be efficiently determined.

A network with n sensors is modeled by a point formation $(\mathbb{G}, p)$, where each sensor corresponds to exactly one vertex of $\mathbb{G}$, and vice versa, with (i, j) being an edge of $\mathbb{G}$ if either i and j are both anchors or the distance between the corresponding sensors is known and $p = \{p_1, \ldots, p_n\}$, where p_i is the position of the sensor corresponding to vertex i. We say that $\mathbb{G}$ is the graph of the network and p is the multipoint of the network. In this work we will be concerned only with networks in the plane. It is known that if the multipoint of a network in $\mathbb{R}^2$ is generic, then the network is localizable if and only if it has at least 3 noncollinear anchors and the graph of the network is globally rigid in $\mathbb{R}^2$ [5]. Since *almost all* multipoints are generic, we will, without loss of generality, restrict our attention to those networks with generic multipoints [4]. In particular, for networks in the plane, this implies no two sensors occupy the same point and no three sensors are collinear in the networks we consider. For such networks, the localizability of the network depends only on the number of its anchors and its graph. Because we are concerned only with networks in the plane, we will refer to graphs that are globally rigid, or rigid, in $\mathbb{R}^2$ as simply globally rigid, or rigid. To avoid trivial and degenerate cases, we will restrict our attention to networks containing four or more sensors.

In the following, let $\mathbb{N}$ be a localizable network of $n > 3$ sensors in the plane labeled 1 through n, and suppose the multipoint of the point formation modeling $\mathbb{N}$ is generic. Let $\mathbb{G} = (\mathcal{V}, \mathcal{E})$ be the graph of $\mathbb{N}$. Since the multipoint of $\mathbb{N}$ is assumed to be generic, we have that $\mathbb{N}$ is localizable if and only if $\mathbb{G}$ is globally rigid in $\mathbb{R}^2$ and $\mathbb{N}$ has at least three anchors. As noted previously, there are a number of efficient algorithms for determining if a graph is globally rigid in $\mathbb{R}^2$ [11, 9, 10]. Hence, it follows that the localizability of $\mathbb{N}$ can also be efficiently determined just by analyzing the graph of $\mathbb{N}$ and counting its anchors. Without loss of generality, suppose that for each $i \in \{1, 2, \ldots, n\}$, vertex i of $\mathbb{G}$ corresponds to sensor i and vice versa. For each $v \in \mathcal{V}$, let $\mathcal{N}(v)$ denote the set consisting of all vertices u where $(u, v) \in \mathcal{E}$, and for each $u \in \mathcal{N}(v)$ write d_{uv} for the distance between sensors u and v.

3. Sequentially localizable networks. Suppose $\mathcal{A}$ is a set of at least three sensors of $\mathbb{N}$ and the vertices corresponding to the sensors in $\mathcal{A}$ induce a complete graph in $\mathbb{G}$, i.e., the distances among *all* pairs of sensors in $\mathcal{A}$ are known. Suppose a position $\pi(a)$ is assigned to each sensor $a \in \mathcal{A}$ such that all known distances among the sensors of $\mathcal{A}$ are satisfied. Since $\mathbb{N}$ is localizable, it is straightforward to show that the positions assigned to the sensors in $\mathcal{A}$ determine a unique position for each of the sensors not in $\mathcal{A}$. In other words, there corresponds exactly one position $\pi(v)$ to each of the sensors $v \in \mathcal{V} - \mathcal{A}$ such that all known intersensor distances are satisfied, i.e., $\|\pi(v) - \pi(u)\| = d_{uv}$ for all $(u, v) \in \mathcal{E}$. We call $\pi(v)$, $v \in \mathcal{V}$, the position of sensor v *relative* to $\mathcal{A}$, and we call $\mathcal{A}$ the set of *proxy anchors* of $\mathbb{N}$. It is easy to show that if sensors labeled a_1, a_2, a_3 are three anchors whose positions are given either by GPS or manual configuration, then the given anchor positions and $\pi(a_1), \pi(a_2), \pi(a_3)$ can be used to compute a Euclidean transformation which maps each $\pi(v)$, $v \in \mathcal{V}$, to the actual position of sensor v.

Let $\mathcal{A}$ denote a set of at least three proxy anchors of $\mathbb{N}$, i.e., $\mathcal{A}$ is any set of three sensors for which all distances among the sensors are given, and each sensor in $\mathcal{A}$ has been assigned a position so that the given distances among them are satisfied. Let $\pi(u)$, $u \in \mathcal{A}$, denote the position assigned to sensor u, and let $\pi(v)$, $v \in \mathcal{V}$, denote the position of sensor v relative to $\mathcal{A}$. For each sensor v and a set $\mathcal{S}$ of points in the plane, we say that $\mathcal{S}$ is a *candidate positions set* of sensor v if $\pi(v) \in \mathcal{S}$. If a candidate positions set consists of a finite number of points, then the set is said to be *finite*. By a *sweep* of $\mathbb{N}$ is meant any sequence $v_1, \ldots, v_n$ obtained by relabeling the n sensors in any way. By a *predecessor* of a sensor in a sweep is meant any other sensor preceding it in the sweep such that the distance between the two sensors is known. The concatenation of a finite number of sweeps in a specific order is a *multiple sweep.* By a *sequential localization algorithm* is meant any localization algorithm which processes the sensors in a network, one by one, in a predetermined sequence in such a way so that the sequence is a multiple sweep and the data for each successive sensor $v \in \mathcal{V} - \mathcal{A}$ in the sequence are either the empty set or a finite candidate positions set for v computed using only the known distances between v and its predecessors, and previously determined data for v and the predecessors of v. The data for sensor $a \in \mathcal{A}$ is assumed to be the singleton candidate positions set consisting of just its assigned position. Clearly, the position of sensor v relative to $\mathcal{A}$ is computed if a candidate positions set consisting of just one element is computed for v. Suppose a singleton candidate positions set has been computed for each sensor of $\mathbb{N}$. If $\mathbb{N}$ has three anchors, then the given positions of the anchors and their computed positions relative to $\mathcal{A}$ can be used to obtain a Euclidean transformation which maps the computed position of each sensor v to the actual position of sensor v. Since $\mathbb{N}$ is localizable, it must have at least three anchors, so $\mathbb{N}$ can be localized by a sequential localization algorithm followed by a Euclidean transformation. For any localizable network, we say the network is *sequentially localizable* if it can be localized by a sequential localization algorithm followed by a Euclidean transformation. Furthermore, we say the network is sequentially localizable in k *sweeps* if the sequence in which the sensors are processed is a multiple sweep, which is the concatenation of k sweeps.

A graph has a *bilateration ordering* if its vertices can be ordered as $v_1, \ldots, v_n$

so that the subgraph induced by v_1, v_2, and v_3 is complete, and each v_i, $i > 3$, is adjacent to at least *two* distinct vertices v_j, $j < i$. As noted previously, to avoid degenerate cases, all networks considered below will be assumed to contain at least four sensors. The following is an easily shown property of the graphs of sequentially localizable networks.

LEMMA 1. *A network is sequentially localizable only if its graph has a bilateration ordering.*

All proofs, unless otherwise stated, are given in the appendix.

A simple and well-known example of a sequential localization algorithm is based on the *trilateration* operation, where the position of each sensor is determined using its distances to three sensors whose positions have already been determined. Trilateration can be applied to any localizable network in the plane possessing a special type of sweep called a "trilateration ordering," which means an ordering $v_1, \dots, v_n$ of the vertices of the network's graph so that the subgraph induced by v_1, v_2, v_3 is complete and each v_i with $i > 3$ is adjacent to at least *three* distinct vertices v_j, $j < i$ [1]. Clearly, a trilateration ordering of a graph is also a bilateration ordering, while a bilateration ordering is not necessarily a trilateration ordering. Graphs with trilateration orderings are known to be globally rigid in $\mathbb{R}^2$ [1]. Suppose $\mathbb{G}$ has a trilateration ordering $v_1, \dots, v_n$. Assign positions $p_{v_1}, p_{v_2}, p_{v_3}$ to sensors v_1, v_2, v_3, respectively, so that their intersensor distances are satisfied, and let v_1, v_2, v_3 be the proxy anchors of $\mathbb{N}$. As noted previously, the multipoint of $\mathbb{N}$ is assumed to be generic, which implies no three sensor positions are collinear. Hence, beginning with v_4, trilateration can be used to determine a *unique* position p_{v_i} for each sensor v_i, $i > 3$, using the given distances between v_i and its predecessors in the ordering, and the computed positions of its predecessors. It is easy to show that the computed position of each sensor is the position of the sensor relative to the proxy anchors. Moreover, the actual sensor positions can be obtained from the computed positions p_v, $v \in \mathcal{V}$, via a Euclidean transformation. Hence, the network can be localized by a sequence of trilateration operations followed by a Euclidean transformation and is therefore sequentially localizable in one sweep. Furthermore, it is straightforward to show that a network's graph must have a trilateration ordering if the network is sequentially localizable in one sweep.

LEMMA 2. *A localizable network is sequentially localizable in one sweep if and only if its graph has a trilateration ordering.*

From Lemma 2, we have that localizable networks whose graphs have trilateration orderings are sequentially localizable; however, as we will show below, the converse need not be true. The central aims of this paper are to explicitly characterize the class of sequentially localizable networks and to present a sequential localization algorithm, called Sweeps, which can localize all sequentially localizable networks. The main result of this paper is the following.

THEOREM 1. *A localizable network is sequentially localizable if and only if the graph of the network has a bilateration ordering. All sequentially localizable networks are localizable by Sweeps.*

The proof of Theorem 1 is given in section 6. Given a network whose multipoint is generic, it is known that the network is localizable if and only if it has three anchors and its graph is globally rigid in $\mathbb{R}^2$ [5]. From Theorem 1, it follows that if a network has three anchors and its graph is globally rigid and has a bilateration ordering, then the network is sequentially localizable and can be localized by Sweeps. In section 3.1, we give a polynomial time algorithm for determining if a graph has a bilateration ordering, and for identifying a bilateration ordering of the graph when the graph has at least one such ordering. As noted previously, there are polynomial time algorithms for determining if a graph is globally rigid in $\mathbb{R}^2$. Hence, it can be efficiently determined if a network is sequentially localizable, and therefore localizable by Sweeps, just by analyzing the graph of the network.

The graph $\mathbb{H}$ as shown in Figure 1(a) can be easily verified to be three connected since there do not exist two vertices whose removal disconnects the graph. Furthermore, it can be shown using the Pebble Game, for example, that there does not exist an edge of $\mathbb{H}$ whose removal results in a graph which is not rigid in $\mathbb{R}^2$ [11]. Hence, $\mathbb{H}$ is globally rigid in $\mathbb{R}^2$ [9, 10]. It follows then that any network with three anchors and a generic multipoint and whose graph is $\mathbb{H}$ must be localizable [5]. Furthermore, $\mathbb{H}$ has a bilateration ordering but no trilateration ordering. Hence, if $\mathbb{H}$ is the graph of a network with three anchors, then it follows from Theorem 1 that the network is sequentially localizable and can be localized by Sweeps. However, since $\mathbb{H}$ does not have a trilateration ordering, the network cannot be localized by only a sequence of trilateration operations followed by a Euclidean transformation. We note that not all localizable networks are sequentially localizable. For example, the graph in Figure 1(b) does not have a bilateration ordering since any bilateration ordering must begin with three vertices, all of which are in either $\{1, 2, 3, 4, 5\}$ or $\{6, 7, 8, 9, 10\}$, and no vertex in $\{1, 2, 3, 4, 5\}$ is adjacent to at least two vertices in $\{6, 7, 8, 9, 10\}$, and vice versa. Furthermore, it can be checked that the graph is also globally rigid in $\mathbb{R}^2$ [11, 9, 10]. So if the graph in Figure 1(b) is the graph of a network with three anchors, then the network is localizable but not sequentially localizable.

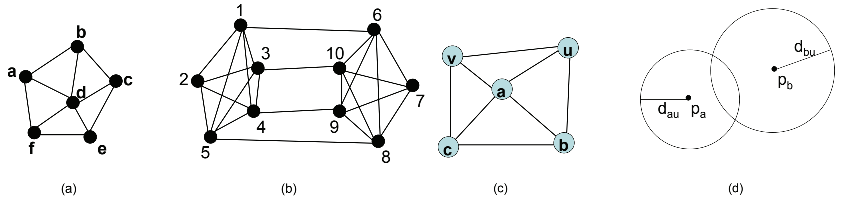


FIG. 1. (a) *A globally rigid graph with a bilateration ordering, i.e., a,d,f,e,c,b, but no trilateration ordering.* (b) *A globally rigid graph without a bilateration ordering.* (c), (d) *Sensors a, b, and c are anchors.*

3.1. Bilateration orderings. As noted in [1] a graph with a trilateration ordering must also be globally rigid in $\mathbb{R}^2$. It is easy to show by example that a graph with a bilateration ordering is not necessarily globally rigid in $\mathbb{R}^2$. However, a graph with a bilateration ordering is necessarily rigid in $\mathbb{R}^2$. More specifically, given any graph which is rigid in $\mathbb{R}^2$, it is known that if a new vertex x is added to the graph by making x adjacent to two or more vertices of the graph, then the resulting graph is again rigid in $\mathbb{R}^2$ [17]. Suppose $\mathbb{G}$ has a bilateration ordering, and let $v_1, v_2, v_3, \ldots, v_n$ be any bilateration ordering of $\mathbb{G}$. Let $\mathbb{G}_i$, $i \in \{3, 4, \ldots, n\}$, denote the graph induced in $\mathbb{G}$ by all vertices v_j where $j \leq i$. Since the complete graph on three vertices is rigid, it follows that $\mathbb{G}_3$ must be rigid. Now suppose $\mathbb{G}_i$ is rigid for some $i \in \{3, \ldots, n-1\}$. Since $\mathbb{G}_{i+1}$ can be obtained from $\mathbb{G}_i$ by making v_{i+1} adjacent to two or more vertices of $\mathbb{G}_i$, it follows that $\mathbb{G}_{i+1}$ must also be rigid. By induction then, $\mathbb{G}_n$, and therefore $\mathbb{G}$, must be rigid in $\mathbb{R}^2$. It is known that if a graph is rigid in $\mathbb{R}^2$, then the graph must also be two connected. Therefore, any graph with a bilateration ordering is rigid in $\mathbb{R}^2$ and two connected.

A graph may have zero, one, or multiple bilateration orderings. If there is no set of three vertices of $\mathbb{G}$ which induce a complete subgraph in $\mathbb{G}$, then $\mathbb{G}$ cannot have a bilateration ordering. Suppose $\mathbb{G}$ has at least one set of three vertices which induce a complete subgraph. In the following we give a polynomial time algorithm for determining if $\mathbb{G}$ has a bilateration ordering, and for identifying a bilateration ordering of $\mathbb{G}$ if there is at least one such ordering. Let x, y, z be any set of three vertices which induce a complete subgraph in $\mathbb{G}$, and let $\mathcal{W}_1 = \{x\}$, $\mathcal{W}_2 = \{y\}$, $\mathcal{W}_3 = \{z\}$. Suppose $\mathcal{W}_i$ for some $i \geq 3$ has been defined. If there exists a vertex u in $\mathcal{V} - \bigcup_{j \leq i} \mathcal{W}_j$ such that u is adjacent to at least two vertices in $\bigcup_{j \leq i} \mathcal{W}_j$, then define $\mathcal{W}_{i+1} = \{u\}$. Otherwise, set $\mathcal{W}_{i+1} = \emptyset$ and stop the algorithm. Let $\mathcal{W}_1, \dots, \mathcal{W}_h$ be the nonempty sets generated by this procedure. Clearly, $h = n$ if and only if there is a bilateration ordering of the graph beginning with x, y, z. If $h = n$, then the ordering obtained by labeling the vertex in $\mathcal{W}_i$, $i \in \{1, \dots, n\}$, as v_i is a bilateration ordering of $\mathbb{G}$. For each $i \in \{1, \dots, h\}$, we have that $|\mathcal{V} - \bigcup_{j \leq i} \mathcal{W}_j| = n - i$, and it takes a number of operations that are linear in n to check if a vertex in $\mathcal{V} - \bigcup_{j \leq i} \mathcal{W}_j$ is adjacent to two vertices in $\bigcup_{j \leq i} \mathcal{W}_j$. Hence, it takes a number of operations that are polynomial in n to determine each of the sets $\mathcal{W}_1, \dots, \mathcal{W}_h$, where h is at most n. The vertex labeling to obtain a bilateration ordering is clearly a linear time procedure. Furthermore, there are at most $\binom{n}{3}$ possible choices for the first three vertices of a bilateration ordering, which implies it can be determined in polynomial time if a graph has a bilateration ordering, and to identify a bilateration ordering if it exists. Since the global rigidity of a graph in $\mathbb{R}^2$ can also be efficiently determined, we conclude that it can be determined in polynomial time if a network is sequentially localizable just by analyzing the graph of the network.

In the following, we describe a class of graphs for which a bilateration ordering can be obtained beginning with *any* two adjacent vertices. A graph is called a *cycle* if its vertices can be relabeled as $c_1, \dots, c_m$, $m \geq 2$, such that c_i is adjacent to c_j if and only if $|i - j| = 1$ or $|i - j| = m - 1$. The *length* of a cycle is the number of edges in the cycle. Let $\mathbb{H}$ be any graph, and let $\mathbb{C}$ be any subgraph of $\mathbb{H}$ such that $\mathbb{C}$ is a cycle. If vertices u and v are adjacent in $\mathbb{H}$, and u and v are nonadjacent vertices in $\mathbb{C}$, then the edge (u, v) is called a *chord* of $\mathbb{C}$. A graph $\mathbb{H}$ is said to be *chordal* if for each subgraph which is also a cycle of length at least four, $\mathbb{H}$ contains at least one chord of the cycle. A chordal graph is not necessarily rigid.

Lemma 3. *Let $\mathbb{H}$ be a rigid and chordal graph with at least four vertices. Then $\mathbb{H}$ has a bilateration ordering, and moreover, for each edge (u, v) in $\mathbb{H}$, there exists a bilateration ordering of $\mathbb{H}$ that begins with vertices u and v.*

Hence, if a graph is rigid and chordal, then the graph has a bilateration ordering, and furthermore, it is particularly easy to determine a bilateration ordering of the graph since any two adjacent vertices must be the first two vertices of some bilateration ordering. An additional simple consequence of Lemma 3 is that any rigid graph which is also chordal must contain a "triangle," i.e., a cycle of length three.

4. The Restricted Sweeps algorithm. In what follows we present a restricted version of the Sweeps algorithm, called Restricted Sweeps, for the class of networks whose graphs have bilateration orderings. The Sweeps algorithm is an extension of this and will be given in section 5.

We begin with an informal description of Restricted Sweeps. A bilateration ordering of the network's graph is first determined, assuming such an ordering exists, and the sensors corresponding to the first three vertices of the ordering are designated the proxy anchors. Positions are assigned to the proxy anchors so that the known distances among them are satisfied. For notational convenience, we assume each vertex of the network's graph has the same label as that of the sensor to which it corresponds. Roughly speaking, the algorithm "sweeps" through the network by processing the sensors sequentially according to the chosen bilateration ordering, beginning with the first sensor in the ordering which is not a proxy anchor. For each sensor which is not a proxy anchor, a finite candidate positions set of the sensor is computed using the known distances from the sensor to its predecessors in the ordering, and the candidate positions sets, or assigned positions, of its predecessors. Recall that a predecessor of a sensor in an ordering is simply any other sensor preceding it in the ordering such that the distance between the two sensors is known. Once the last sensor in the ordering is processed, a candidate positions set will have been computed for each sensor. We call this first "sweep" a *finite candidate positions set generating sweep.* If not every candidate positions set generated by the first sweep is a singleton, then subsequent "refining" sweeps are performed to remove, if possible, elements from each candidate positions set so as to obtain a candidate positions set of fewer elements. To perform a refining sweep, an ordering distinct from the one used to perform the previous sweep is determined, and the sensors are again processed sequentially according to the new ordering. In section 8, we will give a polynomial time algorithm for determining orderings so as to localize the network in as few sweeps as possible by analyzing the graph of the network. At the very least, the new ordering should be such that at least one sensor with a nonsingleton candidate positions set has a predecessor in the ordering. For each sensor v which is not a proxy anchor and whose candidate positions set is not a singleton, the candidate positions sets of v's predecessors in the new ordering, and the known distances between v and its predecessors, are used to identify, if possible, those points in v's candidate positions set which cannot be sensor v's position relative to the proxy anchors. The identified points are removed from the candidate positions set of sensor v to obtain a candidate positions set of fewer elements. We call each sweep after the first sweep a *refining* sweep since the goal of each subsequent sweep is to obtain smaller candidate positions sets.

To illustrate the general idea of a sweep, we use Restricted Sweeps to localize a simple network whose graph is shown in Figure 1(c). For each pair of adjacent vertices i, j, let d_{ij} denote the known distance between sensors labeled i and j. We assume the multipoint of the network is generic, which implies in particular that no three sensor positions are collinear. Vertices a, b, c correspond to the anchors, and vertices u, v correspond to sensors whose positions are to be determined. Let p_a, p_b, and p_c denote the positions of anchors a, b, and c, respectively. It can be efficiently determined that the graph in Figure 1(c) is globally rigid in $\mathbb{R}^2$ [9, 11, 10]. Since the network has three anchors and a generic multipoint, it follows that the network must be localizable.

For the sake of notational convenience, we choose for the first sweep the ordering a, b, c, u, v, in which case the proxy anchors correspond to the actual anchors. The algorithm begins by letting the candidate positions set of each anchor be the singleton set consisting of the anchor's position. Hence, the positions of sensors u and v relative to the anchors a, b, c are simply their actual positions. For a point $p \in \mathbb{R}^2$ and a positive real number r, let $\mathcal{C}(p, r)$ denote the circle with center p and radius r. Since u is the first nonanchor sensor in the chosen ordering, the algorithm proceeds in the first sweep by computing a finite candidate positions set for sensor u, which is just the set of points where the two circles $\mathcal{C}(p_a, d_{au})$ and $\mathcal{C}(p_b, d_{bu})$ intersect. Since no three sensor positions are collinear, it follows that $\mathcal{C}(p_a, d_{au})$ and $\mathcal{C}(p_b, d_{bu})$ must intersect at exactly two points. For instance, if $\mathcal{C}(p_a, d_{au})$ and $\mathcal{C}(p_b, d_{bu})$ are as shown in Figure 1(d), then the two points of intersection comprise the candidate positions set of sensor u. Let $\mathcal{S}(u,1)$ denote the candidate positions set computed for sensor u. Once a finite candidate positions set has been computed for sensor u, Restricted Sweeps proceeds in the first sweep by determining a finite candidate positions set for sensor v as follows. For each point $p \in \mathcal{S}(u,1)$ which is distinct from both p_a and p_c, let $\mathcal{I}(p)$ denote the points in the common intersection of the three circles $\mathcal{C}(p_a, d_{av})$, $\mathcal{C}(p_c, d_{cv})$, and $\mathcal{C}(p, d_{uv})$. The candidate positions set computed for sensor v, denoted $\mathcal{S}(v,1)$, is the union of all $\mathcal{I}(p)$, $p \in \mathcal{S}(u,1)$ where $p \neq p_a$ and $p \neq p_c$. Now we show that $\mathcal{S}(v,1)$ must be a singleton. Let p_v be any point in $\mathcal{S}(v,1)$. Clearly, the distance from p_v to anchors a and c must be d_{av} and d_{cv}, respectively. Furthermore, since p_v is in the intersection of $\mathcal{C}(p_a, d_{av})$, $\mathcal{C}(p_c, d_{cv})$, and $\mathcal{C}(p_u, d_{uv})$ for some $p_u \in \mathcal{S}(u,1)$, we have that the distance between p_u and p_v must be d_{uv}. Note that since $p_u \in \mathcal{S}(u,1)$, it follows that the distance between p_u and anchors a and b must be d_{ua} and d_{ub}, respectively. In other words, for each point $p_v \in \mathcal{S}(v,1)$, there exists a point $p_u \in \mathcal{S}(u,1)$ such that all known intersensor distances are satisfied when sensors v and u are assigned positions p_v and p_u, respectively, and the anchors are simply assigned their given positions. Since the network is localizable, we have that there exists exactly one point corresponding to each nonanchor sensor such that all known intersensor distances are satisfied. Hence, it must be the case that $\mathcal{S}(v,1)$ is a singleton. By definition, the point in $\mathcal{S}(v,1)$ must be the position of sensor v, so the first sweep not only computes a finite candidate positions set for sensor v, but also localizes sensor v since the computed candidate positions set is a singleton.

After the first sweep, a finite candidate positions set will have been determined for both u and v. Since the candidate positions set of sensor u, i.e., $\mathcal{S}(u,1)$, is not a singleton, a second ordering is determined in order to perform a refining sweep. Let the second ordering be a, b, c, v, u. Notice that the ordering has a sensor with a nonsingleton candidate positions set, namely u, that also has at least one predecessor in the ordering. More specifically, the predecessors of sensor u in the second ordering are sensors a, b, and v. The second sweep begins by considering the first vertex in the chosen ordering which has a nonsingleton candidate positions set, which in this case would be sensor u. The Restricted Sweeps algorithm identifies, and removes, points in $\mathcal{S}(u,1)$ which cannot be the position of sensor u as follows. The key observation is that if a point $p \in \mathcal{S}(u,1)$ is the position of sensor u, then for *each* of u's predecessors, there must exist a point in the candidate positions set of the predecessor such that p's distance to that point is the known distance between u and the predecessor. If this is not the case, then the point p can be removed from $\mathcal{S}(u,1)$ to obtain a new candidate positions set of fewer points.

Now we show that the second sweep will remove all but the actual position of sensor u from $\mathcal{S}(u,1)$. First, note that the distances between the actual position of sensor u and the points in the singleton candidate positions sets of a, b, and v must be d_{ua}, d_{ub}, and d_{uv}, respectively. So the actual position of sensor u will not be removed from $\mathcal{S}(u,1)$. Suppose there is a point $q \in \mathcal{S}(u,1)$ such that q is not removed by the second sweep; i.e., the distances between point q and the points in the singleton candidate positions sets of a, b, and v are d_{ua}, d_{ub}, and d_{uv}, respectively. Clearly, if sensor u is assigned point q as its position, and sensors a, b, c, v are assigned their actual positions, then all known intersensor distances are satisfied. Since the network is localizable, we have that there exists exactly one position corresponding to each nonanchor sensor such that all known intersensor distances are satisfied. This implies point q must be the actual position of sensor u, and all other points in $\mathcal{S}(u,1)$ will have been removed from $\mathcal{S}(u,1)$ by the second sweep. Hence, the second sweep localizes sensor u since the candidate positions set of u is a singleton after the second sweep.

Also, since the candidate positions set of each sensor is a singleton after the second sweep, it follows that Restricted Sweeps can localize the network in two sweeps.

4.1. Restricted Sweeps. Suppose the network $\mathbb{N}$ is localizable and the graph of $\mathbb{N}$, i.e., $\mathbb{G}$, has at least one bilateration ordering. We first give the terms and definitions to be used in describing the Restricted Sweeps algorithm. Let $2^{\mathbb{R}^2}$ be the power set of $\mathbb{R}^2$ and write $\mathbb{R}_+$ for the set of positive real numbers. Let $f : 2^{\mathbb{R}^2} \times \mathbb{R}_+ \to 2^{\mathbb{R}^2}$ denote the function $(\mathcal{S}, d) \longmapsto \mathcal{S}'$, where $\mathcal{S}'$ is the set of $p \in \mathbb{R}^2$ such that $\|p - t\| = d$ for some $t \in \mathcal{S}$. If $\mathcal{S} \in 2^{\mathbb{R}^2}$ is not empty, then geometrically $f(\mathcal{S}, d)$ is the union of all points in the plane which lie on circles with the same radius d centered at the points in $\mathcal{S}$. Of course if $\mathcal{S}$ is empty, then so is $f(\mathcal{S}, d)$ and conversely. We will be especially interested in the case when $\mathcal{S}$ is a nonempty "finite set" and $d > 0$, where by *finite set* we mean a set with a finite number of points in $\mathbb{R}^2$. In this case $f(\mathcal{S}, d)$ is simply the union of a finite number of circles in the plane which all have radius d.

Let $\mathbb{S}$ denote the set of all nonempty subsets of $\mathbb{R}^2$ with finitely many elements. Let q be a positive integer no smaller than 2 and write $\mathbb{S}^q$ for the q-fold Cartesian product of $\mathbb{S}$ with itself. Similarly, let $(\mathbb{R}_+)^q$ denote the q-fold Cartesian product of $\mathbb{R}_+$ with itself. Our goal is to define a function $g_q : \mathbb{S}^q \times (\mathbb{R}_+)^q \to 2^{\mathbb{R}^2}$ in such a way so that for each $\{\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_q\} \in \mathbb{S}^q$ and $\{d_1, d_2, \dots, d_q\} \in (\mathbb{R}_+)^q$, $g_q(\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_q, d_1, d_2, \dots, d_q)$ is at most a finite set. Furthermore, we shall require the definition of g_q to be such that whenever there are distinct points $u_i \in \mathcal{S}_i$, $i \in \{1, 2, \dots, q\}$, if $v \in \mathbb{R}^2$ satisfies $\|v - u_i\| = d_i$, $i \in \{1, 2, \dots, q\}$, then v must be a point in $g_q(\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_q, d_1, d_2, \dots, d_q)$. Defining $g_q(\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_q, d_1, d_2, \dots, d_q)$ in the most obvious way as the intersection of the sets $f(\mathcal{S}_i, d_i)$, $i \in \{1, 2, \dots, q\}$, will not be adequate, for it may be the case that the resulting intersection is a continuous circle of points in the plane rather than a finite set. However, a necessary condition for this to occur is that $\bigcap_{j=1}^q \mathcal{S}_j \neq \emptyset$. Hence, let $\mathcal{I} = \bigcap_{j=1}^q \mathcal{S}_j$ and let $\mathcal{X}$ be the intersection of the sets $f(\mathcal{S}_1 \backslash \mathcal{I}, d_1)$ and $f(\mathcal{S}_i, d_i)$, $i \in \{2, \dots, q\}$, which is clearly finite. For each point p in $\mathcal{I}$, let $\mathcal{Y}(p)$ denote the intersection of $f(\{p\}, d_1)$ and $f(\mathcal{S}_i \backslash \{p\}, d_i)$, $i \in \{2, \dots, q\}$, which is again finite. By letting $g_q(\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_q, d_1, d_2, \dots, d_q)$ be the union of $\mathcal{X}$ and $\mathcal{Y}(p)$, $p \in \mathcal{I}$, it is easy to see that g_q satisfies all the aforementioned requirements.

More formally, for $\mathcal{S}_1, \dots, \mathcal{S}_q \in \mathbb{S}$, let $\mathcal{I} = \bigcap_{j=1}^q \mathcal{S}_j$. Let k denote the number of elements in $\mathcal{I}$. If $\mathcal{I}$ is not the empty set, i.e., $k > 0$, then let $p_1, p_2, \dots, p_k$ denote the elements of $\mathcal{I}$. For any set $\mathcal{S} \in \mathbb{S}$, and any subset $\mathcal{T} \subseteq \mathcal{S}$, let $\mathcal{S} \backslash \mathcal{T}$ denote the complement of $\mathcal{T}$ in $\mathcal{S}$. Define the function $g_q : \mathbb{S}^q \times (\mathbb{R}_+)^q \to 2^{\mathbb{R}^2}$ as follows:

$$\begin{aligned} g_q(\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_q, d_1, d_2, \dots, d_q) &= \Big(f(\mathcal{S}_1 \backslash \mathcal{I}, d_1) \cap f(\mathcal{S}_2, d_2) \cap \cdots \cap f(\mathcal{S}_q, d_q) \Big) \\ &\bigcup \left(\bigcup_{i=1}^{k} f(\{p_i\}, d_1) \cap f(\mathcal{S}_2 \backslash \{p_i\}, d_2) \cap \dots\dots \cap f(\mathcal{S}_q \backslash \{p_i\}, d_q) \right). \end{aligned} \tag{1}$$

For $q \geq 2$, it is easy to show that g_q is defined such that for each $\{\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_q\} \in \mathbb{S}^q$ and $\{d_1, d_2, \dots, d_q\} \in (\mathbb{R}_+)^q$, $g_q(\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_q, d_1, d_2, \dots, d_q)$ is at most a finite set. Furthermore, whenever there are distinct points $u_i \in \mathcal{S}_i$, $i \in \{1, 2, \dots, q\}$, if $v \in \mathbb{R}^2$ satisfies $\|v - u_i\| = d_i$, $i \in \{1, 2, \dots, q\}$, then v must be a point in $g_q(\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_q, d_1, d_2, \dots, d_q)$.

Let $[v] = v_1, v_2, v_3, \dots, v_n$ be a bilateration ordering of $\mathbb{G}$. We begin by assigning a point $\pi(i)$ in $\mathbb{R}^2$ to each v_i, $i \in \{1, 2, 3\}$, so that the given distances among the sensors corresponding to v_i, $i \in \{1,2,3\}$, are satisfied. Let the proxy anchors of $\mathbb{N}$ be v_1, v_2, v_3. For each v_i, $i > 3$, let $\pi(v_i)$ denote the position of sensor v_i relative to the proxy anchors. In the following, we will describe an iterative procedure for computing a sequence of candidate positions sets for each $v \in \mathcal{V}$, i.e., $\mathcal{S}(v,1), \mathcal{S}(v,2), \dots, \mathcal{S}(v,i), \dots$.

For $i \in \{4,\dots,n\}$, let $\mathcal{M}(v_i) = \mathcal{N}(v_i) \cap \{v_1, v_2, \dots, v_{i-1}\}$. We denote the cardinality of $\mathcal{M}(v_i)$ by q_i and the elements of $\mathcal{M}(v_i)$ by $u_{i1}, u_{i2}, \dots, u_{iq_i}$. Clearly $q_i \geq 2$ for all $i \in \{4,\dots,n\}$ since $[v]$ is a bilateration ordering. We define the sets $\mathcal{S}(v_i,1)$, $i \in \{1,2,\dots,n\}$, as follows. For $i \in \{1,2,3\}$, let

$$\mathcal{S}(v_i,1) = \{\pi(i)\}, \tag{2}$$

and for $i \in \{4,5,\dots,n\}$, let

$$\mathcal{S}(v_i,1) = g_{q_i}(\mathcal{S}(u_{i1},1), \mathcal{S}(u_{i2},1), \dots, \mathcal{S}(u_{iq_i},1), d_{u_{i1}v_i}, d_{u_{i2}v_i}, \dots, d_{u_{iq_i}v_i}). \tag{3}$$

Suppose $\mathcal{S}(v,k)$, $v \in \mathcal{V}$, have been computed. The sets $\mathcal{S}(v,k+1)$, $v \in \mathcal{V}$, are computed as follows. Let $[x] = x_1, x_2, \dots, x_n$ be an ordering of $\mathcal{V}$, and for $i \in \{1,\dots,n\}$ let $\mathcal{M}(x_i) = \mathcal{N}(x_i) \cap \{x_1, x_2, \dots, x_{i-1}\}$. Note that $[x]$ need not be a bilateration ordering. For $i \in \{1,2,3,\dots,n\}$, if $\mathcal{M}(x_i) = \emptyset$ or $|\mathcal{S}(x_i,k)| = 1$, then let

$$\mathcal{S}(x_i,k+1) = \mathcal{S}(x_i,k). \tag{4}$$

Otherwise, let

$$\mathcal{S}(x_i,k+1) = \mathcal{S}(x_i,k) \bigcap \left(\bigcap_{w \in \mathcal{M}(x_i)} f(\mathcal{S}(w,k+1), d_{wx_i}) \right). \tag{5}$$

4.2. Properties of the Restricted Sweeps algorithm. In the following, we will show that for all $v \in \mathcal{V}$, each $\mathcal{S}(v,i)$ is a finite candidate positions set for v, i.e., $\pi(v) \in \mathcal{S}(v,i)$, and $\mathcal{S}(v,j) \subseteq \mathcal{S}(v,i)$ if $i < j$.

Since $[v]$ is assumed to be a bilateration ordering, each $\mathcal{M}(v_i)$, $i > 3$, has at least two elements and so $q_i \geq 2$. Hence, for $i \in \{4,\dots,n\}$, g_{q_i} is defined, and (3) implies that $\mathcal{S}(v_i,1)$ is a finite set because the image of g_{q_i} consists of only finite sets. Since $\mathcal{S}(v_i,1)$, $i \in \{1,2,3\}$, are also finite sets because of (2), we have that $\mathcal{S}(v,1)$ is a finite set for each $v \in \mathcal{V}$. Note also that $\pi(v_i) \in \mathcal{S}(v_i,1)$, $v_i \in \mathcal{V}$. This is clearly true for $i \in \{1,2,3\}$ because of (2). For any vertex $v \in \mathcal{V}$, an easily verified property of the function f is that if $u \in \mathcal{N}(v)$, and $\mathcal{S}(u)$ is a set for which $\pi(u) \in \mathcal{S}(u)$, then $\pi(v) \in f(\mathcal{S}(u), d_{uv})$. We call this the *position keeping* property of f. The fact that $\pi(v)$, $v \in \mathcal{V}$, are distinct, together with the definition of g_{q_i} and the position keeping property of f, implies that $\pi(v_i) \in \mathcal{S}(v_i,1)$ for $i \in \{4,\dots,n\}$. So each $\mathcal{S}(v,1)$, $v \in \mathcal{V}$, is a finite candidate positions set for sensor v, and we call the computation of $\mathcal{S}(v,1)$, $v \in \mathcal{V}$, a *finite position generating sweep* of $\mathbb{N}$. Suppose for some $k \geq 1$ that $\mathcal{S}(v,k)$, $v \in \mathcal{V}$, is a finite candidate positions set for v, i.e., $\pi(v) \in \mathcal{S}(v,k)$. For each x_i, (5) and (4) imply that $\mathcal{S}(x_i,k+1)$ must be a finite set since $\mathcal{S}(x_i,k)$ is a finite set. The fact that $\pi(x_i) \in \mathcal{S}(x_i,k)$ and the position keeping property of f imply $\pi(x_i) \in \mathcal{S}(x_i,k+1)$ for each x_i. So for each $v \in \mathcal{V}$, $\mathcal{S}(v,k+1)$ is a finite candidate positions set for v; furthermore, it is obvious from (5) and (4) that $\mathcal{S}(v,k+1) \subseteq \mathcal{S}(v,k)$ for all $v \in \mathcal{V}$. It follows from (2), (3), (5), and (4) that each $\mathcal{S}(v,i)$, $v \in \mathcal{V}$, $i \in \{1,\dots,k+1\}$, is computed using $\mathcal{S}(u,i)$, where u is a predecessor of v in the ordering chosen for the ith sweep, and $\mathcal{S}(v, i-1)$ when $i > 1$. By definition then, the Restricted Sweeps algorithm is a sequential localization algorithm.

The preceding shows that if we sweep through the network a finite number of times beginning with a finite position generating sweep, we can generate a sequence of finite candidate positions sets for each $v \in \mathcal{V}$, i.e., $\mathcal{S}(v, 1), \mathcal{S}(v, 2), \dots, \mathcal{S}(v, i), \dots,$ such that $\mathcal{S}(v, 1) \supseteq \mathcal{S}(v, 2) \supseteq \cdots \supseteq \mathcal{S}(v, i) \supseteq \cdots$. As we will show in section 4.3, each $\mathcal{S}(v, 1)$, $v \in \mathcal{V}$, is obtained by solving a sequence of a finite number of quadratic equations, and each $\mathcal{S}(v, i)$, $v \in \mathcal{V}$, $i > 1$, is obtained by computing the distance between a finite number of specified pairs of points. Thus if we can sweep through the network a finite number of times, say k, such that for all $v \in \mathcal{V}$, each $\mathcal{S}(v, k)$ will contain just one element, then that element must be $\pi(v)$. Since $\mathbb{N}$ is localizable, $\mathbb{N}$ must have at least three anchors, so the sensor positions can be obtained from $\pi(v)$, $v \in \mathcal{V}$, via a Euclidean transformation computed from the anchors. In this case, we say the network is localizable by the Restricted Sweeps algorithm *in k sweeps* followed by a Euclidean transformation. In section 8, we will give the graph properties of networks for which we can choose sweep orderings so that the first sweep is a finite position generating sweep and the network is localized in as few sweeps as possible. We will also describe the procedure by which we can efficiently determine the sweep orderings by analyzing the network's graph.

4.3. Quadratic equations. The localization of $\mathbb{N}$ can be formulated mathematically as a system of $|\mathcal{E}|$ simultaneous quadratic equations in $|\mathcal{V}|$ variables:

$$(x_i - x_j)^2 + (y_i - y_j)^2 = d_{ij}^2 \quad \forall\ (i, j) \in \mathcal{E}, \tag{6}$$

where (x_i, y_i) denotes the unknown position of sensor i. In the following we show that the Restricted Sweeps algorithm is equivalent to solving a sequence of a finite number of quadratic equations, where each equation has just one unknown, the solution of which is easily obtained by the well-known quadratic formula, and computing the distance between a finite number of specified pairs of points.

We first consider the computation of $\mathcal{S}(v, 1)$, $v \in \mathcal{V}$. Each $\mathcal{S}(v, 1)$ is defined using the function g_q defined in (1). Since the ordering used for the first sweep must be a bilateration ordering, it must be the case that q is at least 2 in (3). Computing $g_q(\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_q, d_1, d_2, \dots, d_q)$ is equivalent to solving the following system of equations in variables x and y for each collection of q points (a_i, b_i), $i \in \{1, \dots, q\}$, where each $(a_i, b_i) \in \mathcal{S}_i$ and not all q points are identical:

$$(x - a_i)^2 + (y - b_i)^2 = d_i^2, \quad i \in \{1, \dots, q\}. \tag{7}$$

First, consider the case where $q = 2$. The equations in (7) become

$$(x - a_1)^2 + (y - b_1)^2 = d_1^2, \tag{8}$$

$$(x - a_2)^2 + (y - b_2)^2 = d_2^2. \tag{9}$$

Equations (8) and (9) are satisfied by the coordinates of the points of intersection, if any, of the two circles with radii d_1 and d_2, and centered at (a_1, b_1) and (a_2, b_2), respectively. Since (a_1, b_1) and (a_2, b_2) are assumed to be nonidentical, the coordinates of at most two points in the plane can satisfy (8) and (9). See Figure 2 for the three cases where the two circles intersect at two, one, and zero points respectively. Equations (8) and (9) can be rewritten as one quadratic equation in one variable in the obvious way. Since (a_i, b_i), $i \in \{1, \dots, q\}$, are assumed to be distinct, it must be

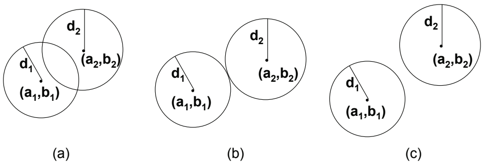


FIG. 2. (a) *Two intersection points.* (b) *One intersection point.* (c) *Zero intersection points.*

the case that either $a_1 - a_2 \neq 0$ or $b_1 - b_2 \neq 0$. Without loss of generality, suppose the latter is true. By subtracting (9) from (8), the quadratic terms cancel, and we get

$$y = \frac{d_1^2 - d_2^2 - (b_1^2 - b_2^2 + a_1^2 - a_2^2) - (-2a_1 + 2a_2)x}{-2b_1 + 2b_2}. \tag{10}$$

Hence, (8) can be rewritten as a quadratic equation of just the variable x:

$$(x - a_1)^2 + \left(\frac{d_1^2 - d_2^2 - (b_1^2 - b_2^2 + a_1^2 - a_2^2) - (-2a_1 + 2a_2)x}{-2b_1 + 2b_2} - b_1\right)^2 = d_1^2. \tag{11}$$

Obviously, if (x, y) satisfies (8) and (9), then x must satisfy (11). On the other hand, suppose x satisfies (11) and y satisfies (10); then x and y must also satisfy (8). So, if we let $P_1(x, y) = (x - a_1)^2 + (y - b_1)^2 - d_1^2$, $P_2(x, y) = (x - a_2)^2 + (y - b_2)^2 - d_2^2$, and $P_3(x, y) = P_1(x, y) - P_2(x, y)$, then x, y satisfy $P_1(x, y) = 0$ and $P_3(x, y) = 0$. Since $P_3(x, y) = P_1(x, y) - P_2(x, y)$, this implies $P_2(x, y) = 0$, which implies that (9) is also satisfied by x and y. Therefore, x and y satisfy (8) and (9) if and only if they also satisfy (11) and (10). Hence, for the case where $q = 2$, solving for x, y which satisfy (8) and (9) reduces to solving a quadratic equation in x and then solving for y via substitution. Since we are interested only in points in the real plane whose coordinates satisfy (8) and (9), any complex solutions to (8) and (9) are discarded. Clearly, when $q > 2$, solving for x, y which satisfy (7) can be similarly reduced to solving a quadratic equation in x and then solving for y via substitution. Furthermore, it is not difficult to show that when $q > 2$ the solution to (7) can be obtained by just solving a linear system of equations. Since each $\mathcal{S}_j$, $j \in \{1, \ldots, q\}$, is a finite set, it follows that computing $\mathcal{S}(v, 1)$, $v \in \mathcal{V}$, in Restricted Sweeps is equivalent to solving a sequence of a finite number of polynomial equations, each in one variable and each with degree at most two, the solution of which is easily obtained by the quadratic formula.

Now consider the computation of $\mathcal{S}(v, k)$, $v \in \mathcal{V}$, for $k > 1$. Let $\mathcal{M}(v)$ denote the vertices adjacent to v which also precede v in the ordering chosen for the kth sweep. If $\mathcal{M}(v) = \emptyset$, then $\mathcal{S}(v, k) = \mathcal{S}(v, k - 1)$, so suppose $\mathcal{M}(v)$ is nonempty, and let $u_1, \ldots, u_m$, $m \geq 1$, denote the elements of $\mathcal{M}(v)$. When $\mathcal{M}(v)$ is nonempty, $\mathcal{S}(v, k)$ is computed using (5). It follows from (5) that $\mathcal{S}(v, k)$ is obtained by removing all points p from $\mathcal{S}(v, k-1)$ for which there does not exist points $p_1 \in \mathcal{S}(u_1, k), \ldots, p_m \in \mathcal{S}(u_m, k)$ such that $\|p - p_i\| = d_{u_i v}$ for each $i \in \{1, \ldots, m\}$. Hence, (5) consists of computing the distances between pairs of points in $\mathcal{S}(v, k - 1)$ and $\mathcal{S}(u_i, k)$, $i \in \{1, \ldots, m\}$. From this, we conclude that Restricted Sweeps is equivalent to solving a sequence of a finite number of quadratic equations, where the solution of each equation is easily obtained by the well-known quadratic formula, and computing the distance between a finite number of specified pairs of points.

5. The Sweeps algorithm. An extension to the Restricted Sweeps algorithm was proposed in [7]. The Sweeps algorithm to be presented in what follows is based upon the extension to the Restricted Sweeps algorithm which we now describe.

Like the Restricted Sweeps algorithm, Sweeps is a localization algorithm for the class of networks whose graphs have bilateration orderings. As in Restricted Sweeps, the Sweeps algorithm "sweeps" through the network according to a predetermined bilateration ordering of the sensors and computes a finite candidate positions set for each sensor using the candidate positions sets of its predecessors and known distances. The key difference in Sweeps is that a "subassignment" function is associated with each point in the candidate positions set computed for a sensor. We illustrate this using a simple example. We first define an *assignment* of $\mathbb{N}$ to be any function $\alpha : \mathcal{V} \rightarrow \mathbb{R}^2$. By a *subassignment* of $\mathbb{N}$ is meant any function that is the restriction of an assignment to a nonempty subset of $\mathcal{V}$. Suppose u, v, w, x is a subsequence of the ordering chosen for the first sweep, i.e., the finite candidate positions set generating sweep, and suppose v and w are each adjacent to both u and x, as shown in Figure 3(a). Let $\mathcal{S}(u, 1)$, $\mathcal{S}(v, 1)$, and $\mathcal{S}(w, 1)$ denote the candidate positions sets of u, v, w, respectively, computed in the first sweep by Restricted Sweeps. Since u is a predecessor of both v and w in the ordering, $\mathcal{S}(u, 1)$ is used in the computations of both $\mathcal{S}(v, 1)$ and $\mathcal{S}(w, 1)$. More specifically, suppose v has predecessors u and u'. From (3), we have that each point $p_v \in \mathcal{S}(v, 1)$ is obtained by computing the intersection of circles centered at distinct points p_u and $p_{u'}$ for some $p_u \in \mathcal{S}(u, 1)$ and $p_{u'} \in \mathcal{S}(u', 1)$. Hence, p_v can be considered a candidate position of sensor v under the assumption that sensors u and u' are positioned at p_u and $p_{u'}$, respectively. A graphical illustration of this is shown in Figure 3(b).

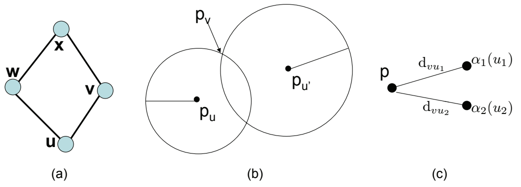


FIG. 3.

In the Restricted Sweeps algorithm, the candidate positions set of sensor v contains no "record" of the fact that p_v was computed assuming u is positioned at p_u and u' is positioned at $p_{u'}$. The Sweeps algorithm extends Restricted Sweeps by using a subassignment to keep track of the fact that p_v was computed assuming sensors u and u' are positioned at p_u and $p_{u'}$, respectively. So a subassignment β is associated with p_v, where the domain of β contains v, u, u' and $\beta(v) = p_v$, $\beta(u) = p_u$, and $\beta(u') = p_{u'}$. More generally, for each sensor v and each point p in the candidate positions set of v, the assumed position of each sensor whose candidate positions set was either directly or indirectly used in computing p is kept track of via a subassignment function. In reference to Figure 3(a), suppose $p_v \in \mathcal{S}(v, 1)$ is computed assuming sensor u is positioned at p_u, and $q_w \in \mathcal{S}(w, 1)$ is computed assuming sensor u is positioned at q_u, where $q_u \neq p_u$. Since both v and w are predecessors of x, we have that the candidate positions sets of both v and w are used in computing the candidate positions set of x. For the sake of this example, suppose that the only predecessors of x are v and w. In Restricted Sweeps, p_v and q_w would be used in computing the candidate positions set of x. More specifically, if the circle centered at p_v with radius d_{vx} and the circle centered at q_w with radius d_{wx} intersect at one or more points, then each of those points would be an element in the candidate positions set of x. In Sweeps, however, p_v and q_w would *not* be used in computing the candidate positions set of sensor x

because the two points were computed assuming different positions for sensor u. For certain networks, the candidate positions sets generated by Sweeps contain significantly fewer elements than those generated by Restricted Sweeps. And as we will see in section 7, the computational complexity of localizing a network by Sweeps, or Restricted Sweeps, is entirely dependent on the number of elements in the generated candidate positions sets.

5.1. Sweeps. Suppose the network $\mathbb{N}$ is localizable and the graph of $\mathbb{N}$, i.e., $\mathbb{G}$, has at least one bilateration ordering. We first give the terms and definitions to be used in describing the Sweeps algorithm. An assignment α is *consistent* if $\|\alpha(u)-\alpha(v)\| = d_{uv}$ for all $(u, v) \in \mathcal{E}$. Let $\mathcal{D}(\alpha)$ denote the domain of a subassignment α. Two subassignments α and β are said to be *consistent with each other*, and we write $\alpha \sim \beta$, if there does not exist $u \in \mathcal{D}(\alpha) \cap \mathcal{D}(\beta)$ such that $\alpha(u) \neq \beta(u)$. For $p \in \mathbb{R}^2$ and a positive real number r, let $\mathcal{C}(p, r)$ denote the circle of radius r centered at p. Let $\alpha_1, \dots, \alpha_k$ be a collection of $k \geq 1$ pairwise consistent subassignments, i.e., $\alpha_i \sim \alpha_j$ for all $i, j \in \{1, \dots, k\}$, and define $u_k(\alpha_1, \dots, \alpha_k)$ as the subassignment with domain $\bigcup_{i\in\{1,\dots,k\}} \mathcal{D}(\alpha_i)$ whose restriction to $\mathcal{D}(\alpha_i)$ is equal to α_i for each $i \in \{1, \dots, k\}$.

Consider a collection of $k \geq 2$ pairwise consistent subassignments $\alpha_1, \dots, \alpha_k$. Suppose there are vertices $v \in \mathcal{V}$ and $u_i \in \mathcal{D}(\alpha_i)$, $i \in \{1, \dots, k\}$, such that $(v, u_i) \in \mathcal{E}$ for all $i \in \{1, \dots, k\}$, and v is not an element of the domain of any α_i. If there is a point p whose distance to each $\alpha_i(u_i)$, $i \in \{1, \dots, k\}$, is d_{vu_i}, then roughly speaking, p is a candidate position for sensor v assuming each sensor u_i, $i \in \{1, \dots, k\}$, is positioned at $\alpha_i(u_i)$. More generally, p can be viewed as a candidate position for sensor v assuming each sensor $u \in \bigcup_{i\in\{1,\dots,k\}} \mathcal{D}(\alpha_i)$ is positioned at $\alpha(u)$, where $\alpha = u_k(\alpha_1, \dots, \alpha_k)$. We aim to define a set $\mathcal{M}(\alpha_1, \dots, \alpha_k, v, u_1, \dots, u_k)$ with the goal of keeping track of the candidate positions of sensor v assuming sensors u_i, $i \in \{1, \dots, k\}$, are positioned at $\alpha_i(u_i)$, $i \in \{1, \dots, k\}$, respectively. Since sensor positions are assumed to be distinct, we shall be interested only in the case where $\alpha_i(u_i)$, $i \in \{1, \dots, k\}$, are distinct. See Figure 3(c) for an illustration of the case when $k = 2$. To keep track of the fact that p is a candidate position for sensor v assuming each sensor u_i, $i \in \{1, \dots, k\}$, is positioned at $\alpha_i(u_i)$, define the subassignment β^p with domain $\{v\} \cup \bigcup_{i\in\{1,\dots,k\}} \mathcal{D}(\alpha_i)$ such that $\beta^p(v) = p$ and $\beta^p(u_i) = \alpha_i(u_i)$ for each $i \in \{1, \dots, k\}$:

$$\beta^p(v) = p, \quad \beta^p(u) = \zeta(u) \quad \forall\ u \in \bigcup_{i\in\{1,\dots,k\}} \mathcal{D}(\alpha_i), \tag{12}$$

where $\zeta = u_k(\alpha_1, \dots, \alpha_k)$. Let $\mathcal{M}(\alpha_1, \dots, \alpha_k, v, u_1, \dots, u_k)$ denote the set of all such β^p. More formally, if $\bigcap_{j\in\{1,\dots,k\}} \mathcal{C}(\alpha_j(u_j), d_{vu_j}) = \emptyset$, or $\alpha_i(u_i) = \alpha_j(u_j)$ for some $i, j \in \{1, \dots, k\}$, $i \neq j$, then let $\mathcal{M}(\alpha_1, \dots, \alpha_k, v, u_1, \dots, u_k) = \emptyset$. Otherwise, since $k \geq 2$, it is easy to see that $\bigcap_{j\in\{1,\dots,k\}} \mathcal{C}(\alpha_j(u_j), d_{vu_j})$ is a set consisting of at most q points in $\mathbb{R}^2$, where q is at most 2. Let the points be denoted by $p_1, \dots, p_q$, and let

$$\mathcal{M}(\alpha_1, \dots, \alpha_k, v, u_1, \dots, u_k) = \{\beta^{p_1}, \dots, \beta^{p_q}\}. \tag{13}$$

In the Sweeps algorithm, a sequence of finite sets of subassignments $\mathcal{S}(v, 1), \dots, \mathcal{S}(v, j)$ is computed for each $v \in \mathcal{V}$, where for each $i \in \{1, \dots, j\}$, v is in the domain of each subassignment in $\mathcal{S}(v, i)$, and $\{\beta(v) \mid \beta \in \mathcal{S}(v, i)\}$ is a finite candidate positions set for v.

Let $[v] = v_1, v_2, v_3, \dots, v_n$ be a bilateration ordering of $\mathbb{G}$. We begin by assigning a point $\pi(v_i)$ in $\mathbb{R}^2$ to each v_i, $i \in \{1, 2, 3\}$, so that the known distances among the sensors corresponding to v_i, $i \in \{1, 2, 3\}$, are satisfied. Let the proxy anchors of $\mathbb{N}$ be

v_1, v_2, v_3. For each v_i, $i > 3$, let $\pi(v_i)$ denote the position of sensor v_i relative to the proxy anchors. For v_i, $i \in \{1, 2, 3\}$, let α_i be the subassignment with domain $\{v_i\}$, where $\alpha_i(v_i) = \pi(v_i)$. For $i \in \{1, 2, 3\}$, let $\mathcal{S}(v_i, 1)$ be defined as

$$\mathcal{S}(v_i, 1) = \{\alpha_i\}, \quad i \in \{1, 2, 3\}. \tag{14}$$

The sets $\mathcal{S}(v_i, 1)$, $i > 3$, are computed iteratively as follows. For v_i, $i > 3$, let $\mathcal{M}(v_i) = \mathcal{N}(v_i) \cap \{v_1, \dots, v_{i-1}\}$. Since $[v]$ is a bilateration ordering, each $\mathcal{M}(v_i)$, $i > 3$, must be a set of at least two elements. Let $u_1, \dots, u_m$ be the elements of $\mathcal{M}(v_i)$. In order to compute $\mathcal{S}(v_i, 1)$, we consider each collection of pairwise consistent subassignments $\alpha_j \in \mathcal{S}(u_j, 1)$, $j \in \{1, \dots, m\}$. Suppose $\mathcal{M}(\alpha_1, \alpha_2, \dots, \alpha_m, v_i, u_1, u_2, \dots, u_m) \neq \emptyset$, and let $\beta \in \mathcal{M}(\alpha_1, \alpha_2, \dots, \alpha_m, v_i, u_1, u_2, \dots, u_m)$. From (12), we have that $v_i, u_1, \dots, u_m \in \mathcal{D}(\beta)$, $\beta(u_j) = \alpha_j(u_j)$ for all $j \in \{1, \dots, m\}$, and $\beta(v_i)$ is a candidate position of sensor v_i assuming that each sensor $u \in \mathcal{D}(\beta)$ is positioned at $\beta(u)$, and more specifically, that each sensor u_j, $j \in \{1, \dots, m\}$, is positioned at $\alpha_j(u_j)$. The set $\mathcal{S}(v_i, 1)$ is intended to be the set of all such subassignments β, where $\beta \in \mathcal{M}(\alpha_1, \alpha_2, \dots, \alpha_m, v_i, u_1, u_2, \dots, u_m)$ and $\alpha_j \in \mathcal{S}(u_j, 1)$, $j \in \{1, \dots, m\}$ is a collection of pairwise consistent subassignments. Hence, $\mathcal{S}(v_i, 1)$ is defined as

$$\mathcal{S}(v_i, 1) = \bigcup_{\alpha_j \in \mathcal{S}(u_j,1)\ \forall\ j \in \{1,\dots,m\} \text{ and } \alpha_j \sim \alpha_k\ \forall\ j,k \in \{1,\dots,m\}} \mathcal{M}(\alpha_1, \dots, \alpha_m, v_i, u_1, \dots, u_m). \tag{15}$$

Note that since $|\mathcal{M}(v_i)| \geq 2$ for each v_i, where $i > 3$, it follows that each $\mathcal{S}(v, 1)$ consists of a finite number of elements.

Suppose for some $k \geq 1$ that $\mathcal{S}(u, k)$, $u \in \mathcal{V}$, have been computed, and that each $\mathcal{S}(u, k)$ consists of a finite number of elements. Let $u_1, \dots, u_n$, denoted $[u]$, be any ordering of the vertices such that the first three vertices of $[u]$ coincide with the proxy anchors: $u_1 = v_1$, $u_2 = v_2$, $u_3 = v_3$. Note that $[u]$ is not required to be a bilateration ordering. Once the ordering $[u]$ is selected, the sets $\mathcal{S}(u, k+1)$, $u \in \mathcal{V}$, are computed iteratively as follows. For $i \in \{1, 2, 3\}$, let $\mathcal{S}(u_i, k+1) = \mathcal{S}(u_i, k)$. For $i \in \{4, \dots, n\}$, let $\mathcal{M}(u_i) = \mathcal{N}(u_i) \cap \{u_1, \dots, u_{i-1}\}$, and let $\mathcal{S}(u_i, k+1) = \mathcal{S}(u_i, k)$ if $\mathcal{M}(u_i) = \emptyset$. If $\mathcal{M}(u_i)$ is nonempty, then let $w_1, \dots, w_m$ be the elements of $\mathcal{M}(u_i)$. For notational convenience, let w_0 denote u_i. Suppose α_0 is a subassignment in $\mathcal{S}(u_i, k)$ for which there exists a collection of subassignments $\alpha_j \in \mathcal{S}(w_j, k+1)$, $j \in \{1, \dots, m\}$, such that $\alpha_0, \alpha_1, \dots, \alpha_m$ are pairwise consistent and $\|\alpha_0(u_i) - \alpha_j(w_j)\| = d_{u_i w_j}$ for *all* w_j, $j \in \{1, \dots, m\}$. In this case, $\alpha_0(u_i)$ can be considered a candidate position for sensor u_i assuming that each w_j, $j \in \{1, \dots, m\}$, is positioned at $\alpha_j(w_j)$, and more generally, that each $w \in \mathcal{D}(\alpha_j)$, $j \in \{1, \dots, m\}$, is positioned at $\alpha_j(w)$. Hence, if α_0 is "augmented" to a subassignment α, where $\alpha = u_{m+1}(\alpha_0, \alpha_1, \dots, \alpha_m)$, then $\alpha(u_i) = \alpha_0(u_i)$ and $\alpha(w_j) = \alpha_j(w_j)$ for all $j \in \{1, \dots, m\}$, and $\alpha(w) = \alpha_j(w)$ for each $w \in \mathcal{D}(\alpha_j)$, $j \in \{1, \dots, m\}$. Roughly speaking, $\mathcal{S}(u_i, k+1)$ is the set of subassignments obtained from $\mathcal{S}(u_i, k)$ by "augmenting" each such subassignment α_0 to $u_{m+1}(\alpha_0, \alpha_1, \dots, \alpha_m)$. Now suppose β is a subassignment in $\mathcal{S}(u_i, k)$ for which there does *not* exist some collection of subassignments $\beta_j \in \mathcal{S}(w_j, k+1)$, $j \in \{1, \dots, m\}$, such that $\beta, \beta_1, \dots, \beta_m$ are pairwise consistent and $\|\beta(u_i) - \beta_j(w_j)\| = d_{u_i w_j}$ for *all* w_j, $j \in \{1, \dots, m\}$. It is straightforward to show that $\beta(u_i)$ cannot be the position of sensor u_i relative to the proxy anchors, and so β is not used to define any subassignment in $\mathcal{S}(u_i, k+1)$. Roughly speaking, $\beta(u_i)$ is removed from consideration as a candidate position for sensor u_i. More formally, $\mathcal{S}(u_i, k+1)$ is defined as

$$\begin{aligned}(16)\qquad \mathcal{S}(u_i, k+1) = \Bigg\{ & u_{m+1}(\alpha_0, \alpha_1, \ldots, \alpha_m) \mid \alpha_0 \in \mathcal{S}(u_i, k), \\ & \alpha_j \in \mathcal{S}(w_j, k+1) \quad \forall\, j \in \{1, \ldots, m\}, \\ & \alpha_h \sim \alpha_j, \quad \alpha_h(w_h) \neq \alpha_j(w_j) \quad \forall\, h, j \in \{0, 1, \ldots, m\}, \\ (17)\qquad & \alpha_0(u_i) \in \bigcap_{j \in \{1, \ldots, m\}} \mathcal{C}(\alpha_j(w_j), d_{u_i w_j}) \Bigg\}.\end{aligned}$$

Since each $\mathcal{S}(v, k)$, $v \in \mathcal{V}$, consists of a finite number of elements, it follows from (17) that $\mathcal{S}(v, k+1)$ must also consist of a finite number of elements.

By the same argument as that used in section 4.3, it follows that Sweeps is equivalent to solving a sequence of a finite number of quadratic equations, where each equation has just one unknown, the solution of which is easily obtained by the well-known quadratic formula, and computing the distance between a finite number of specified pairs of points.

5.2. Properties of Sweeps.

As noted previously, each of the sets computed by the Sweeps algorithm consists of a finite number of subassignments. In the following, we give some additional properties of these sets.

LEMMA 4. *Let w be any vertex of $\mathbb{G}$. If vertices u, v are adjacent in $\mathbb{G}$, and $u, v \in \mathcal{D}(\beta)$ for some $\beta \in \mathcal{S}(w, 1)$, then $\|\beta(u) - \beta(v)\| = d_{uv}$.*

From (17), we have that each subassignment $\alpha \in \mathcal{S}(v, 2)$, $v \in \mathcal{V}$, is a subassignment "augmented" from some subassignment $\hat{\alpha} \in \mathcal{S}(v, 1)$, i.e., $\alpha = u_{m+1}(\hat{\alpha}, \alpha_1, \ldots, \alpha_m)$. From this and Lemma 4 we can show the following.

LEMMA 5. *Let w be any vertex of $\mathbb{G}$. If vertices u, v are adjacent in $\mathbb{G}$, and $u, v \in \mathcal{D}(\beta)$ for some $\beta \in \mathcal{S}(w, 2)$, then $\|\beta(u) - \beta(v)\| = d_{uv}$.*

Recall that $v_1, \ldots, v_n$ was the ordering used to compute $\mathcal{S}(v, 1)$, $v \in \mathcal{V}$, and v_1, v_2, v_3 are the proxy anchors of $\mathbb{N}$ whose assigned positions are $\pi(v_1)$, $\pi(v_2)$, and $\pi(v_3)$, respectively. Since $\mathbb{N}$ is localizable, there exists exactly one consistent assignment $\bar{\alpha}$ of $\mathbb{N}$, where $\bar{\alpha}(v_i) = \pi(v_i)$ for each $i \in \{1, 2, 3\}$. Furthermore, for each $v \in \mathcal{V}$, $\bar{\alpha}(v)$ is the position of sensor v relative to the proxy anchors. Suppose $\mathcal{S}(v, 1), \mathcal{S}(v, 2), \ldots, \mathcal{S}(v, k)$ are computed for each $v \in \mathcal{V}$.

LEMMA 6. *For each $v \in \mathcal{V}$ and $i \in \{1, \ldots, k\}$, there is a $\beta \in \mathcal{S}(v, i)$ which is the restriction of $\bar{\alpha}$ to the domain of β and $v \in \mathcal{D}(\beta)$.*

Each $\mathcal{S}(v_i, k)$, $k \geq 1$, is computed using sets $\mathcal{S}(v_j, k)$, where v_j is a predecessor of v_i in the kth chosen ordering, and $\mathcal{S}(v_i, k-1)$ when $k > 1$. Recall that the sensors of $\mathbb{N}$ are labeled $1, \ldots, n$ and $\mathcal{V} = \{1, \ldots, n\}$. Each subassignment β may be represented as a sequence of n points, where the ith point in the sequence is $\beta(i)$ if $i \in \mathcal{D}(\beta)$ and is $\emptyset$ otherwise. Hence, if $\beta_1, \ldots, \beta_m$ are m subassignments, where subassignment β_i is represented by the sequence $p_{i1}, \ldots, p_{in}$, then the set consisting of subassignments $\beta_1, \ldots, \beta_m$ can be represented by the set of points $\{p_{11}, \ldots, p_{1n}, p_{21}, \ldots, p_{2n}, \ldots, p_{m1}, \ldots, p_{mn}\}$. From Lemma 6 we have that for each sensor v and each computed ith sweep, the set $\{\beta(v) \mid \beta \in \mathcal{S}(v, i)\}$ is a finite candidate positions set for v. By definition then, Sweeps is a sequential localization algorithm.

6. Graphical properties of networks localizable by Sweeps.

In the following, we show that the necessary condition for a localizable network to be sequentially localizable is also a sufficient condition for the network to be localizable by the Sweeps algorithm. More specifically, we show that all localizable networks whose graphs have bilateration orderings can be localized by computing $\mathcal{S}(v,k)$, $v \in \mathcal{V}$, where $k \leq 2$, with the Sweeps algorithm, and we give an efficient algorithm for determining the sensor ordering of each sweep.

Let $\mathcal{A}$ denote any set of three vertices in $\mathbb{G}$ which induce a complete graph in $\mathbb{G}$. Let $\mathbb{H}_1, \mathbb{H}_2, \ldots, \mathbb{H}_c$ denote the maximally connected components of the subgraph of $\mathbb{G}$ induced by vertices in $\mathcal{V} - \mathcal{A}$. The following is a consequence of the assumption that $\mathbb{N}$ is localizable.

Lemma 7. *For each $i \in \{1,\ldots,c\}$, the graph induced in $\mathbb{G}$ by $\mathcal{A}$ and the vertices of $\mathbb{H}_i$ is globally rigid in $\mathbb{R}^2$.*

Let $\mathbb{H}$ be $\mathbb{H}_i$ for any $i \in \{1,\ldots,c\}$, and let u be any vertex of $\mathbb{H}$. In the following we construct a partition of the vertex set of $\mathbb{H}$. Let $\mathcal{N}_0(u) = \{u\}$, and let $\mathcal{N}_1(u)$ denote the set of vertices in $\mathcal{V} - \mathcal{A}$ adjacent to u. Suppose for some integer $i \geq 1$, $\mathcal{N}_j(u)$, $j \in \{0,1,\ldots,i\}$, have been determined. Let $\mathcal{N}_{i+1}(u)$ denote the set of vertices $w \in \mathcal{V} - \mathcal{A}$, where $w \notin \bigcup_{j\in\{0,\ldots,i\}} \mathcal{N}_j(u)$ and w is adjacent to a vertex in $\mathcal{N}_i(u)$. Since there are a finite number of vertices, there can be only a finite number of sets generated this way. Suppose we have $h+1$ sets generated this way: $\mathcal{N}_0(u), \mathcal{N}_1(u), \ldots, \mathcal{N}_h(u)$. It is straightforward to show that the sets $\mathcal{N}_i(u)$, $i \in \{0,1,\ldots,h\}$, make up a partition of the vertices of $\mathbb{H}$. We call $\mathcal{N}_i(u)$, $i \in \{0,1,\ldots,h\}$, a *vertex partition* of $\mathbb{H}$. Let n' denote the number of vertices in $\mathbb{H}$. Select any n' elements of $\{|\mathcal{A}|+1,\ldots,n\}$, and order them as $i_1, i_2, \ldots, i_{n'}$ so that $i_1 < i_2 < \cdots < i_{n'}$. Assign indices 1 to $|\mathcal{A}|$ to vertices in $\mathcal{A}$ in any manner, and assign index $i_{n'}$ to vertex u. Assign the remaining indices i_j, $j \in \{1,2,\ldots,n'-1\}$, to vertices in $\mathcal{N}_1, \ldots, \mathcal{N}_h$ beginning with $\mathcal{N}_1$ and $n'-1$; i.e., assign indices $i_{n'-1}$ to $i_{n'-|\mathcal{N}_1(u)|}$ to the vertices in $\mathcal{N}_1(u)$ in any manner, assign indices $i_{n'-|\mathcal{N}_1(u)|-1}$ to $i_{n'-|\mathcal{N}_1(u)|-|\mathcal{N}_2(u)|}$ to the vertices in $\mathcal{N}_2(u)$ in any manner, and so on. We call this ordering a *complete ordering* of the vertices of $\mathbb{H}$ *with respect to u and $\mathcal{A}$*, or just a *complete ordering* of the vertices of $\mathbb{H}$ *with respect to $\mathcal{A}$*.

For each $i \in \{1,\ldots,c\}$, let u_i be any vertex in $\mathbb{H}_i$. Since the vertex sets of $\mathbb{H}_i$, $i \in \{1,\ldots,c\}$, are pairwise disjoint, we can construct an ordering of $\mathcal{V}$ that is a complete ordering of $\mathbb{H}_i$ with respect to u_i and $\mathcal{A}$ for all $i \in \{1,\ldots,c\}$. We call this a *complete ordering* of $\mathbb{G}$ *with respect to $u_1,\ldots,u_c$ and $\mathcal{A}$*, or just a *complete ordering* of $\mathbb{G}$ *with respect to $\mathcal{A}$*.

Let $v_1, v_2, v_3, \ldots, v_n$ be any bilateration ordering of $\mathbb{G}$, and suppose $\mathcal{S}(v,1)$, $v \in \mathcal{V}$, are computed using this ordering. This implies that sensors corresponding to v_1, v_2, v_3 make up the set of proxy anchors, and v_1, v_2, v_3 induce a complete subgraph in $\mathbb{G}$. Let $\mathcal{A} = \{v_1, v_2, v_3\}$, and let $\pi(v_1)$, $\pi(v_2)$, and $\pi(v_3)$ be the positions assigned to the proxy anchors v_1, v_2, v_3, respectively. Since $\mathbb{N}$ is localizable, there is exactly one consistent assignment $\bar{\alpha}$ of $\mathbb{N}$ such that $\bar{\alpha}(v_i) = \pi(v_i)$, $i \in \{1,2,3\}$. As noted previously, the actual sensor positions can be obtained from $\bar{\alpha}(v)$, $v \in \mathcal{V}$, via a Euclidean transformation computed using anchor positions. For $i \in \{1,\ldots,c\}$, let u_i be any vertex in $\mathbb{H}_i$.

Lemma 8. *Suppose the ordering used to compute the second sweep, i.e., $\mathcal{S}(v,2)$, $v \in \mathcal{V}$, is a complete ordering of $\mathbb{G}$ with respect to $u_1,\ldots,u_c$ and $\{v_1,v_2,v_3\}$. For each $i \in \{1,\ldots,c\}$, and all $\alpha \in \mathcal{S}(u_i,2)$, $\mathcal{D}(\alpha)$ is the union of $\{v_1,v_2,v_3\}$ and the vertex set of $\mathbb{H}_i$.*

The following is a consequence of Lemmas 5, 6, 7, and 8.

Lemma 9. *Suppose the ordering used to compute the second sweep, i.e., $\mathcal{S}(v,2)$, $v \in \mathcal{V}$, is a complete ordering of $\mathbb{G}$ with respect to $u_1,\ldots,u_c$ and $\{v_1,v_2,v_3\}$. For each $i \in \{1,\ldots,c\}$, $\mathcal{S}(u_i,2)$ is a singleton, and the subassignment in $\mathcal{S}(u_i,2)$ is the restriction of $\bar{\alpha}$ to the union of $\{v_1,v_2,v_3\}$ and the vertex set of $\mathbb{H}_i$.*

If the ordering used to compute the second sweep, i.e., $\mathcal{S}(v,2)$, $v \in \mathcal{V}$, is a complete ordering of $\mathbb{G}$ with respect to $u_1, \dots, u_c$ and $\{v_1, v_2, v_3\}$, then Lemma 9 implies that each $\mathcal{S}(u_i, 2)$, $i \in \{1, \dots, c\}$, consists of exactly one subassignment α_i, which is the restriction of $\bar{\alpha}$ to the union of $\mathcal{A}$ and the vertex set of $\mathbb{H}_i$. Each sensor v which is not a proxy anchor must correspond to a vertex in exactly one of the $\mathbb{H}_i$, $i \in \{1, \dots, c\}$. If sensor v corresponds to a vertex in $\mathbb{H}_i$, then the position of sensor v relative to the proxy anchors, i.e., $\bar{\alpha}(v)$, is given by $\alpha_i(v)$.[1] We have just shown the following.

LEMMA 10. *If $\mathbb{N}$ is localizable and its graph has a bilateration ordering, then $\mathbb{N}$ can be localized by computing two sweeps of the Sweeps algorithm followed by a Euclidean transformation. The ordering of the first sweep is any bilateration ordering $v_1, v_2, v_3, \dots, v_n$, and the ordering of the second sweep is a complete ordering of $\mathbb{G}$ with respect to $\{v_1, v_2, v_3\}$.*

Now we give the proof for Theorem 1. From Lemma 1, we have that a localizable network is sequentially localizable only if its graph has a bilateration ordering. Lemma 10 implies that Sweeps can localize all sequentially localizable networks since a sequentially localizable network's graph must have a bilateration ordering. Furthermore, since Sweeps is a sequential localization algorithm, Lemma 10 implies that a localizable network is sequentially localizable if its graph has a bilateration ordering. Hence, Lemmas 10 and 1 imply a localizable network is sequentially localizable if and only if its graph has a bilateration ordering.

In [7], it was shown via extensive simulations that Sweeps is practically feasible on uniformly random networks of 250 sensors with connectivity modeled by unit disk graphs despite having a worst case computational complexity that is exponential in the number of sensors. In section 7, we give the graph properties of some networks which can be efficiently localized using Sweeps.

7. Efficiently localizable networks.

Consider a class of networks such that for each positive integer i, there is a network in the class with at least i sensors. We say that the class of networks is efficiently localizable by Sweeps (or Restricted Sweeps) if there is a constant c such that each network in the class can be localized by Sweeps (or Restricted Sweeps) in a number of operations that is at most n^c, where n is the number of the network's sensors. The computational complexity of localizing $\mathbb{N}$ by Sweeps, or Restricted Sweeps, is entirely dependent upon the number of elements in the sets $\mathcal{S}(v,1)$, $v \in \mathcal{V}$. More specifically, let $\mathcal{M}(v)$ denote the vertices preceding v and also adjacent to v in the ordering chosen for the first sweep. In both the Sweeps and Restricted Sweeps algorithm, the number of operations necessary to compute $\mathcal{S}(v,1)$, $v \in \mathcal{V}$, is equal to $C\prod_{u\in\mathcal{M}(v)} |\mathcal{S}(u,1)|$, where C is a constant that is independent of the number of sensors in $\mathbb{N}$. In the following, we give a graphical characterization for when a network is efficiently localizable by Sweeps and Restricted Sweeps. We emphasize that this is not a complete characterization of all such efficiently localizable networks. However, the general techniques used here can be used to determine additional efficiently localizable networks.

Suppose the graph of $\mathbb{N}$, namely $\mathbb{G}$, has a trilateration ordering, and that the ordering chosen for the first sweep is a trilateration ordering $v_1, \dots, v_n$. It is easy to see that $\mathcal{S}(v,1)$ is a singleton and $\prod_{u\in\mathcal{M}(v)} |\mathcal{S}(u,1)| = 1$ for all $v \in \mathcal{V}$. Hence, the class of networks whose graphs have trilateration orderings is obviously efficiently localizable by Sweeps. The key property of the trilateration ordering which makes $\mathbb{N}$ efficiently localizable by Sweeps is that for *all* $i \in \{4, \dots, n\}$, the graph induced in $\mathbb{G}$

[1] By a slight modification to the Sweeps algorithm, a singleton candidate positions set can be obtained for each sensor; however, we omit this step since it is unnecessary.

by vertex v_i, and all the vertices v_j, where $j < i$, is globally rigid. We now "relax" this property to define a "superbilateration" ordering. A graph with $n \geq 4$ vertices has a *superbilateration* ordering $v_1, v_2, v_3, \ldots, v_n$ if the graph contains a subgraph with the same vertex set which can be constructed inductively as follows beginning with the complete graph on three vertices labeled v_1, v_2, v_3. Suppose the graph being constructed already contains vertices $v_1, \ldots, v_i$, $i \geq 3$. If $i+1$ is even, then v_{i+1} is added to the graph by making v_{i+1} adjacent to at least three vertices v_j where $j < i+1$. Otherwise, if $i+1$ is odd, then v_{i+1} can be added to the graph in one of two ways, the first of which is to make v_{i+1} adjacent to at least three vertices v_j where $j < i+1$. Or, v_{i+1} can be added to the graph by making v_{i+1} adjacent to distinct vertices v_i, v_k, v_j, where v_k is adjacent to v_i, and removing the edge between v_i and v_k.

A 1-*extension* on a graph is the operation whereby two adjacent vertices of the graph are first selected, say vertices u and v, and a new vertex w is added to the graph by making w adjacent to vertices u, v, and x, where x is distinct from both u and v, and removing the edge between u and v [8]. An *edge-addition* on a graph is the operation whereby two nonadjacent vertices are made adjacent by insertion of a new edge. In [10], it was shown that the graph resulting from an edge-addition or 1-extension operation on any globally rigid graph of four or more vertices is again globally rigid. From this, it follows that any graph with a superbilateration ordering $v_1, \ldots, v_n$ is necessarily globally rigid. Furthermore, for each $i > 3$, where either i is equal to n or i is odd, the graph induced by all vertices v_j, $j \leq i$, is globally rigid. Clearly, a trilateration ordering is automatically a superbilateration ordering. It is easy to show by example that the converse need not be true. Suppose $\mathbb{G}$ has a superbilateration ordering $v_1, \ldots, v_n$. Let v_1, v_2, v_3 be the proxy anchors of $\mathbb{N}$. For $i > 3$, and where i is either odd or equal to n, let $\mathbb{N}_i$ denote the subnetwork consisting of all sensors corresponding to vertices v_j, $j \leq i$. Each subnetwork $\mathbb{N}_i$ can be efficiently localized, relative to the proxy anchors, by Sweeps assuming the positions of all sensors in $\mathbb{N}_i$ which are also in some $\mathbb{N}_j$, $j < i$, are known. Hence, the entire network can be localized in a number of operations polynomial in the number of sensors by using Sweeps to localize each of the subnetworks in sequence beginning with $\mathbb{N}_5$. Generally speaking, suppose a localizable network contains subnetworks $\mathbb{N}_1, \ldots, \mathbb{N}_m$ so that each subnetwork $\mathbb{N}_i$ is efficiently localizable by Sweeps (or Restricted Sweeps) assuming the position of each sensor in $\mathbb{N}_i$ which is also in some $\mathbb{N}_j$, $j < i$, is known. Then the entire network is efficiently localizable by localizing the subnetworks $\mathbb{N}_1, \ldots, \mathbb{N}_m$ in sequence, provided each sensor of $\mathbb{N}$ is in some $\mathbb{N}_i$, $i \in \{1, \ldots, m\}$.

8. Graphical properties of networks localizable by Restricted Sweeps. In this section, we will give sufficient conditions on the graphs of localizable networks for which we can choose sweep orderings so that the network is localized in as few sweeps as possible by the Restricted Sweeps algorithm. First, consider the case where $\mathbb{N}$'s graph $\mathbb{G}$ has a trilateration ordering $v_1, \ldots, v_n$, and suppose this is the ordering chosen for the first sweep. Since we assume that the multipoints of the networks we consider are generic, it follows that no three sensor positions of $\mathbb{N}$ are collinear. Hence, each $\mathcal{S}(v, 1)$, $v \in \mathcal{V}$, as computed by the first sweep of the Restricted Sweeps algorithm, is a singleton. This and Lemma 2 imply that a network with three or more anchors can be localized by Restricted Sweeps in one sweep followed by a Euclidean transformation if and only if its graph has a trilateration ordering.

8.1. Networks with partially acyclic graphs. In the following, we show that if $\mathbb{G}$ is "partially acyclic," then $\mathbb{G}$ must have a bilateration ordering and $\mathbb{N}$ is localizable by Restricted Sweeps in two sweeps plus a Euclidean transformation. For any subset $\mathcal{W}$ of $\mathcal{V}$, let $\mathbb{G}(\mathcal{W})$ denote the graph induced in $\mathbb{G}$ by vertices in $\mathcal{W}$. For any nonempty subset $\mathcal{W}$ of $\mathcal{V}$, we say that $\mathbb{G}$ is *partially acyclic with respect to* $\mathcal{W}$, or just *partially acyclic*, if $\mathbb{G}(\mathcal{W})$ is a complete graph and $\mathbb{G}(\mathcal{V}-\mathcal{W})$ is acyclic. Suppose $\mathbb{G}$ is partially acyclic with respect to $\mathcal{W}$, and that each vertex in $\mathcal{V}-\mathcal{W}$ has degree at least three in $\mathbb{G}$. In the following, we will construct a bilateration ordering of $\mathbb{G}$. We first note that a necessary condition for a graph with four or more vertices to be globally rigid in $\mathbb{R}^2$ is that each of its vertices must have degree at least three [9]. Since $\mathbb{N}$ is localizable and contains at least four sensors, it follows that each vertex of $\mathbb{G}$ must have degree at least three. The graph in Figure 1(a) is globally rigid and partially acyclic with respect to any three mutually adjacent vertices of the graph. Additional globally rigid graphs which are also partially acyclic can be constructed using the edge-addition and 1-extension operations beginning with the complete graph on four vertices [10].

Let $\mathbb{H}$ denote a maximally connected component of $\mathbb{G}(\mathcal{V}-\mathcal{W})$, and let r denote any vertex of $\mathbb{H}$. Let $[r]$ denote any complete ordering of the vertex set of $\mathbb{H}$ with respect to r and $\mathcal{W}$, and let $\mathcal{N}_0(r), \mathcal{N}_1(r), \dots, \mathcal{N}_h(r)$ denote the vertex partition used to construct the ordering $[r]$. We now show that $[r]$ is a bilateration ordering of the graph induced in $\mathbb{G}$ by the vertices of $\mathbb{H}$ and the vertices in $\mathcal{W}$. Let v be any vertex of $\mathbb{H}$ and suppose $v \in \mathcal{N}_i(r)$ for some $i \in \{0, 1, \dots, h\}$. First suppose $i = 0$, in which case v must equal r. Suppose r is adjacent to $c < 2$ vertices of $\mathcal{W}$. Since $\mathbb{H}$ is a maximally connected component of $\mathbb{G}(\mathcal{V}-\mathcal{W})$, and r has degree at least three in $\mathbb{G}$, we have that r must be adjacent to at least $3 - c > 0$ vertices in $\mathbb{H}$, which implies r must be adjacent to at least three vertices preceding it in the ordering $[r]$. Now suppose $i > 0$. This implies v is adjacent to at least one vertex in $\mathcal{N}_{i-1}(r)$. Moreover, v is adjacent to exactly one vertex in $\mathcal{N}_{i-1}(r)$, for if v is adjacent to two vertices in $\mathcal{N}_{i-1}(r)$, then $\mathbb{H}$ is not acyclic, which implies $\mathbb{G}(\mathcal{V}-\mathcal{W})$ is not acyclic. Similarly, if v is adjacent to a vertex in $\mathcal{N}_i(r)$, then that would again imply $\mathbb{H}$ is not acyclic. Since v has degree three in $\mathbb{G}$, v must be adjacent to at least two vertices in $\mathcal{N}_{i+1}(r) \cup \mathcal{W}$. Since the vertices in $\mathcal{N}_{i+1}(r) \cup \mathcal{W}$ all precede v in $[r]$, it follows that v is adjacent to at least two vertices preceding it in the ordering $[r]$. Now we show that the first three vertices of $[r]$ induce a complete graph. Let x be any vertex in $\mathcal{N}_h(r)$. Since $\mathbb{H}$ is acyclic, x can be adjacent to exactly one vertex in $\mathbb{H}$. Also, since x has degree at least three in $\mathbb{G}$, it follows that x must be adjacent to at least two vertices in $\mathcal{W}$ and so $|\mathcal{W}| \geq 2$. Furthermore, since the vertices in $\mathcal{N}_h(r)$ precede all other vertices in $\mathbb{H}$ in the ordering $[r]$, it follows that $v_3 \in \mathcal{W} \cup \mathcal{N}_h(r)$. Hence, the first three vertices of $[r]$ induce a complete graph, and $[r]$ must therefore be a bilateration ordering. Let $\mathcal{V}(\mathbb{H})$ denote the vertex set of $\mathbb{H}$. We have just shown the following.

LEMMA 11. *If $\mathbb{G}$ is partially acyclic with respect to some $\mathcal{W} \subseteq \mathcal{V}$, and each vertex in $\mathcal{V}-\mathcal{W}$ has degree at least three in $\mathbb{G}$, then any complete ordering of a maximally connected component $\mathbb{H}$ of $\mathbb{G}(\mathcal{V}-\mathcal{W})$ with respect to $\mathcal{W}$ is also a bilateration ordering of the graph $\mathbb{G}(\mathcal{V}(\mathbb{H}) \cup \mathcal{W})$.*

Let $\mathbb{H}_1, \dots, \mathbb{H}_c$ denote the maximally connected components of $\mathbb{G}(\mathcal{V}-\mathcal{W})$. For each $\mathbb{H}_i$, let $v_{i1}, v_{i2}, v_{i3}, \dots, v_{ik}$ be any complete ordering of $\mathbb{H}_i$ with respect to $\mathcal{W}$. This implies that the first $|\mathcal{W}|$ vertices of each of the orderings must be the vertices of $\mathcal{W}$, i.e., $\{v_{i1}, \dots, v_{i|\mathcal{W}|}\} = \mathcal{W}$ for all $i \in \{1, \dots, c\}$. Let $w_1, \dots, w_{|\mathcal{W}|}$ denote the vertices of $\mathcal{W}$. From Lemma 11, we have that each of the orderings $v_{i1}, v_{i2}, v_{i3}, \dots, v_{ik}$ is a bilateration ordering. Therefore, the ordering obtained by concatenating $w_1, \dots, w_{|\mathcal{W}|}$ and $v_{i(|\mathcal{W}|+1)}, \dots, v_{ik}$ for all $i \in \{1, \dots, c\}$, i.e., $w_1, \dots, w_{|\mathcal{W}|}, v_{1(|\mathcal{W}|+1)}, \dots, v_{1k}, \dots, v_{i(|\mathcal{W}|+1)}, \dots, v_{ik}, \dots, v_{c(|\mathcal{W}|+1)}, \dots, v_{ck}$ must be a bilateration ordering, and it is a *complete ordering* of $\mathbb{G}$ with respect to $\mathcal{W}$. We have just shown the following.

LEMMA 12. *If $\mathbb{G}$ is partially acyclic with respect to some $\mathcal{W} \subseteq \mathcal{V}$, and each vertex in $\mathcal{V} - \mathcal{W}$ has degree at least three in $\mathbb{G}$, then $\mathbb{G}$ has a bilateration ordering, and any complete ordering of $\mathbb{G}$ with respect to $\mathcal{W}$ is a bilateration ordering.*

Remark 1. It is known that a necessary condition for a graph with at least four vertices to be globally rigid in $\mathbb{R}^2$ is that the graph must be three connected, i.e., $\mathbb{G}(\mathcal{V} - \mathcal{V}')$ is connected if $|\mathcal{V}'| \leq 2$. Hence, if $\mathbb{G}$ is partially acyclic with respect to $\mathcal{W}$, and $\mathbb{G}(\mathcal{V} - \mathcal{W})$ has more than one connected component, then $|\mathcal{W}| \geq 3$. Since $\mathbb{G}$ is globally rigid, Lemma 7 can be used to show that for each $i \in \{1, \ldots, c\}$, the graph $\mathbb{G}(\mathcal{W} \cup \mathcal{V}_i)$, where $\mathcal{V}_i$ is the vertex set of $\mathbb{H}_i$, must also be globally rigid.

Our main result for networks with partially acyclic graphs is the following.

THEOREM 2. *A localizable network with graph $\mathbb{G}$ is localizable by Restricted Sweeps in two sweeps plus a Euclidean transformation if $\mathbb{G}$ is partially acyclic with respect to some $\mathcal{W} \subseteq \mathcal{V}$. The ordering of the finite position generating sweep is $[v] = v_1, v_2, v_3, \ldots, v_n$, where $[v]$ is a complete ordering of $\mathbb{G}$ with respect to $\mathcal{W}$, and the ordering of the second sweep is $v_1, v_2, v_3, v_n, v_{n-1}, \ldots, v_4$.*

A globally rigid graph in $\mathbb{R}^2$ is said to be *minimally* globally rigid in $\mathbb{R}^2$ if no edge can be removed from the graph without causing the graph to no longer be globally rigid in $\mathbb{R}^2$. A number of globally rigid graphs in $\mathbb{R}^2$ that are partially acyclic with respect to some $\mathcal{W} \subseteq \mathcal{V}$, where $|\mathcal{W}| \geq 3$, are also minimally globally rigid in $\mathbb{R}^2$. Hence, Theorem 2 implies that Restricted Sweeps can localize certain networks with just enough edges in their graphs to ensure localizability. For any $i > 3$, let $\mathbb{W}_i$ denote the graph whose vertices can be labeled as $w_0, w_1, \ldots, w_i$ such that w_0 is adjacent to all other vertices, and vertices $w_1, \ldots, w_i$ induce a cycle in the graph. Any such $\mathbb{W}_i$, $i > 3$, is called a *wheel graph.* It is known that wheel graphs are minimally globally rigid, and it is straightforward to show that any wheel graph is partially acyclic with respect to any three vertices which are mutually adjacent. Hence, any network with three or more anchors and whose graph is a wheel graph is localizable by Restricted Sweeps in two sweeps plus a Euclidean transformation. One can show by example that globally rigid graphs which are also partially acyclic are not limited to just wheel graphs.

Let $\mathcal{N}_T$ denote the class of networks whose graphs have a trilateration ordering, and let $\mathcal{N}_P$ denote the class of networks whose graphs are globally rigid in $\mathbb{R}^2$ and partially acyclic. It is not difficult to show that $\mathcal{N}_T$ and $\mathcal{N}_P$ are not disjoint, $\mathcal{N}_T \nsubseteq \mathcal{N}_P$ and $\mathcal{N}_P \nsubseteq \mathcal{N}_T$. For example, networks with wheel graphs are in $\mathcal{N}_P$ but not $\mathcal{N}_T$, and any network whose graph has a trilateration ordering $v_1, \ldots, v_n$, where $n > 5$ and each v_i, $i > 3$, is adjacent to vertices v_{i-1}, v_{i-2}, and v_{i-3}, is in $\mathcal{N}_T$ but not $\mathcal{N}_P$.

8.2. Networks with ring squared graphs. Many practical networks are such that the distance between two sensors is known if the sensors are within a prescribed sensing radius of each other. Suppose $\bar{\mathbb{N}}$ is such a network and has at least three anchors, and let $\bar{\mathbb{G}}$ be its graph. Define a *ring graph* with ordering $v_1, \ldots, v_n$ as a graph whose vertices can be labeled as $v_1, \ldots, v_n$ so that each vertex v_i, $1 < i < n$, is adjacent to vertices v_{i-1} and v_{i+1}, and vertex v_1 is adjacent to vertex v_n.

LEMMA 13. *If $\bar{\mathbb{G}}$ is a ring graph with ordering $v_1, \ldots, v_n$, then $\bar{\mathbb{N}}$ is localizable in two sweeps plus a Euclidean transformation after doubling the sensing radius of each sensor. The ordering of the first sweep is $v_1, \ldots, v_n$, and the ordering of the second sweep is $v_1, v_2, v_3, v_n, v_{n-1}, \ldots, v_4$.*

Let $\bar{\mathcal{V}}$ and $\bar{\mathcal{E}}$ denote the vertex set and edge set of $\bar{\mathbb{G}}$, respectively. The *second power* of $\bar{\mathbb{G}}$, written $\bar{\mathbb{G}}^2$, is the graph with vertex set $\bar{\mathcal{V}}$ and edge set $\bar{\mathcal{E}} \cup \bar{\mathcal{E}}^2$, where $(i, j) \in \bar{\mathcal{E}}^2$ in the case when $i, j \in \bar{\mathcal{V}}$ and there exists $k \in \bar{\mathcal{V}}$ such that $(i, k), (k, j) \in \bar{\mathcal{E}}$.

A graph is *edge 2-connected* if there exists two paths with no edge in common between each pair of vertices. It is known that the second power of an edge 2-connected graph is globally rigid in $\mathbb{R}^2$ [1]. An important consequence of this and Theorem 13 is that if the graph of a network is edge 2-connected with at least three anchor vertices, and the network is such that the distance between two sensors is known if the sensors are within sensing radius, then the network is sequentially localizable after doubling the sensing radius of all the sensors [1].

9. Conclusion. In this work, we presented Sweeps, a sequential localization algorithm which consists of solving a sequence of a finite number of quadratic equations, and determining the distances between specified pairs of points. We identified the graph properties of all networks which can be localized by Sweeps, as well as the graph properties of some networks which can be efficiently localized by Sweeps. The worst case computational complexity of Sweeps is exponential. However, extensive experimental evaluations on uniformly random networks modeled by unit disk graphs indicate that Sweeps is practically much more efficient [7]. Part of our future work will be to analyze the average case computational complexity of Sweeps. Additionally, the necessary and sufficient condition for a localizable network to be localizable by Sweeps is that the graph of the network has a bilateration ordering. Extensive simulations on uniformly random networks modeled by unit disk graphs suggest that the gap between localizable and sequentially localizable networks is not large [7]. A question that is of interest is if there exists a threshold such that a graph is globally rigid *and* has a bilateration ordering when the average degree of the graph passes the threshold. In [12], a trilateration-based localization algorithm was proposed for networks with inaccurate distance measurements in which sensors are assigned an estimated position only when the estimated position can be provably bounded to be within some known range of the actual sensor position. A similar concept was employed in adapting the Sweeps algorithm for the case of inaccurate distance measurements [6] in that each estimated sensor position can be guaranteed to be within a known distance of the actual sensor position. As part of future research, we aim to fine tune and improve the Sweeps algorithm adapted for inaccurate distance measurements.

A key aspect of wireless sensor networks is that each sensor can interact with only a subset of the sensors in the network. Hence, Sweeps and Restricted Sweeps are proposed on the assumption that the distances between each sensor and only some of the sensors in a network are known. Although the computations in Sweeps and Restricted Sweeps are currently envisioned as being carried out on a central computer, we note that this does not necessarily contradict the distributed nature of a wireless sensor network. For example, in a sensor network deployed for environment monitoring, quantities measured by a sensor, i.e., chemical emissions, and transmitted to a base station, make sense only in the context of the sensor's position. The distance measurements taken by each sensor to, say, nearby sensors can be transmitted to the base station along with whatever quantities the sensor was deployed to monitor. The base station can then run a localization algorithm using the intersensor distance measurements, and thus associate a position with each measured quantity. An important part of our future research will be to design a fully distributed version of Sweeps.

10. Appendix.

Proof of Lemma 1. Suppose $\mathbb{N}$ is sequentially localizable. For each sensor v, let $k(v)$ denote the sweep in which a finite candidate positions set was computed. Order the sensors as $v_1, \ldots, v_n$ so that v_i precedes v_j, i.e., $i < j$, if either $k(v_i) < k(v_j)$ or $k(v_i) = k(v_j)$ and v_i is a predecessor of v_j in the $k(v_i)$th sweep. Consider any v_i

which is not a proxy anchor. First, suppose no distance is known between v_i and any sensor v_j, where $j < i$. This implies that when sensor v_i is processed, there is no known distance between v_i and a sensor whose candidate positions set has already been determined. Hence, there is no data with which to compute a finite candidate positions set for v_i. Now suppose the distance between v_i and exactly one other sensor v_j, $j < i$, is known. This implies that when v_i is processed, its distance to exactly one sensor with a finite candidate positions set is known. By definition of a sequential localization algorithm, a sensor for which a finite candidate positions set has not been computed does not have any position information associated with it. Hence, since sensor positions are distinct, a finite candidate positions set of v_i cannot be determined when just its distance to a single sensor with an already computed candidate positions set is known. We have just shown that if v_i is not a proxy anchor, then v_i must be adjacent to at least two v_j, where $j < i$. This implies that v_1 and v_2 must be proxy anchors, and so $v_1, \ldots, v_n$ is a bilateration ordering. $\square$

Proof of Lemma 2. The "if" direction has already been shown in section 3. The "only if" direction is a straightforward consequence of the following. Given a sensor v, and its distances to k sensors $u_1, \ldots, u_k$ with known positions where no three sensors in $\{v, u_1, \ldots, u_k\}$ are collinear, there exists exactly one position for sensor v such that its distances to all k sensors are satisfied if and only if $k \geq 3$. $\square$

Proof of Lemma 3. Let $\mathcal{V}$ denote the vertex set of $\mathbb{G}$, and let u be any vertex in $\mathcal{V}$. Note that u must be adjacent to at least one other vertex in $\mathbb{G}$ since $\mathbb{G}$ is rigid and therefore connected. Let v be any vertex adjacent to u. Let $u_1, u_2, \ldots, u_m$ be any ordering of $m \leq |\mathcal{V}|$ vertices such that $u_1 = u$, $u_2 = v$, and each u_i, $i \geq 3$, is adjacent to at least two vertices u_k, $k < i$. Moreover, suppose there does not exist any vertex $w \in \mathcal{V} - \{u_1, u_2, \ldots, u_m\}$ which is adjacent to two or more vertices in $\{u_1, u_2, \ldots, u_m\}$. Let $\mathbb{B}$ denote the graph induced in $\mathbb{G}$ by $\{u_1 = u, u_2 = v, \ldots, u_m\}$. Note that $\mathbb{B}$ contains at least two vertices, namely, u and v. In the following we will show that $\mathcal{V} - \{u_1, u_2, \ldots, u_m\} \neq \emptyset$ is a contradiction to the assumption that $\mathbb{G}$ is chordal. So, suppose $\mathcal{V} - \{u_1, u_2, \ldots, u_m\} \neq \emptyset$. Let $\mathbb{F}$ denote a maximally connected component of the graph induced by vertices not in $\mathbb{B}$, i.e., $\mathcal{V} - \{u_1, u_2, \ldots, u_m\}$. Note that $\mathbb{F}$ has at least one vertex since $\mathcal{V} - \{u_1, u_2, \ldots, u_m\} \neq \emptyset$. For an edge incident on vertices a and b, we say that the edge is *from* $\mathbb{B}$ to $\mathbb{F}$ if a is in $\mathbb{B}$ and b is in $\mathbb{F}$. Since $\mathbb{G}$ is rigid, and $\mathbb{F}$ contains at least one vertex, there must be at least two edges e_1 and e_2 from $\mathbb{B}$ to $\mathbb{F}$. Let $e_1, e_2, \ldots, e_c$, $c \geq 2$, denote all the edges from $\mathbb{B}$ to $\mathbb{F}$.

A vertex in $\mathbb{F}$ can be incident on at most one edge from $\mathbb{B}$ to $\mathbb{F}$. For if a vertex w in $\mathbb{F}$ is incident on two edges from $\mathbb{B}$ to $\mathbb{F}$, then obviously, $w \in \mathcal{V} - \{u_1, u_2, \ldots, u_m\}$ and w is adjacent to two or more vertices in $\{u_1, u_2, \ldots, u_m\}$, which contradicts our assumption that there does not exist any vertex $w \in \mathcal{V} - \{u_1, u_2, \ldots, u_m\}$ which is adjacent to two or more vertices in $\{u_1, u_2, \ldots, u_m\}$. Suppose e_1 is incident on vertex z in $\mathbb{B}$, and that all the edges from $\mathbb{B}$ to $\mathbb{F}$ are incident on z. By removing z from $\mathbb{G}$, $\mathbb{G}$ is disconnected since this removes all edges from $\mathbb{B}$ to $\mathbb{F}$, and $\mathbb{F}$ is a maximally connected component of the graph induced by vertices not in $\mathbb{B}$. But $\mathbb{G}$ is rigid, which means it is at least two connected, and therefore it requires the removal of at least two vertices to disconnect $\mathbb{G}$. Hence, there must exist at least one edge from $\mathbb{B}$ to $\mathbb{F}$ which is not incident on z. So there must exist two edges from $\mathbb{B}$ to $\mathbb{F}$ such that the edges are incident on distinct vertices in $\mathbb{B}$. Let e_i and e_j denote two such edges. Also from the above, we have that e_i and e_j are incident on distinct vertices in $\mathbb{F}$. Hence, there exist distinct vertices $b, b' \in \mathbb{B}$ and $f, f' \in \mathbb{F}$ such that b is adjacent to f and b' is adjacent to f'. Since $\mathbb{B}$ is connected, there is a path in $\mathbb{B}$ from b to b'. Let this path be denoted $b_0 = b, b_1, b_2, \ldots, b_B = b'$. Since $\mathbb{F}$ is connected, there is a path in $\mathbb{F}$

from f to f'. Let this path be denoted $f_0 = f, f_1, f_2, \ldots, f_F = f'$.

Let L be the smallest positive integer in $\{1, 2, \ldots, F\}$ such that f_L is adjacent to some vertex in $\{b_1, b_2, \ldots, b_B\}$. Note that such an L must exist since $f' = f_F$ is adjacent to $b' = b_B$. Let $\bar{L}$ be such that $b_{\bar{L}}$ is the vertex in $\{b_1, b_2, \ldots, b_B\}$ to which f_L is adjacent. Note that $\bar{L} > 0$ and $L > 0$. Let T be the largest integer less than L such that f_T is adjacent to some vertex in $\{b_0, b_1, \ldots, b_{\bar{L}-1}\}$. Such a T must exist since $f = f_0$ is adjacent to $b = b_0$, and as noted above, $L, \bar{L} > 0$. Let $\bar{T}$ be such that $b_{\bar{T}}$ is the vertex to which f_T is adjacent in $\{b_0, b_1, \ldots, b_{\bar{L}-1}\}$. By construction, the subgraph of $\mathbb{G}$ with vertices $f_T, f_{T+1}, \ldots, f_L, b_{\bar{L}}, b_{\bar{L}-1}, \ldots, b_{\bar{T}}$ and edges $(f_T, f_{T+1}), \ldots, (f_{L-1}, f_L), (f_L, b_{\bar{L}}), (b_{\bar{L}-1}, b_{\bar{L}-2}), \ldots, (b_{\bar{T}-1}, b_{\bar{T}}), (b_{\bar{T}}, f_T)$ is a cycle. Let this cycle be denoted $\mathbb{C}$. Note that $\mathbb{C}$ contains at least four vertices and so is a cycle of length at least four. Since $\mathbb{G}$ is chordal, there must exist an edge in $\mathbb{G}$ that is also a chord of $\mathbb{C}$. Since f_T and f_L can each be incident upon only one edge from $\mathbb{B}$ to $\mathbb{F}$, we have that any chord of $\mathbb{C}$ that is also an edge from $\mathbb{B}$ to $\mathbb{F}$ must be incident upon f_M, where $T < M < L$. Suppose there is such a vertex f_M. Since L is the smallest positive integer in $\{1, 2, \ldots, F\}$ such that f_L is adjacent to some vertex in $\{b_1, b_2, \ldots, b_B\}$, it follows that f_M must be adjacent to b_0 since $M < L$. But this is a contradiction since $M > T$ and T is the largest integer less than L such that f_T is adjacent to some vertex in $\{b_0, b_1, \ldots, b_{\bar{L}-1}\}$. Hence, any chord of $\mathbb{C}$ can only contain vertices which are either both in $\mathbb{B}$ or both in $\mathbb{F}$. Since $\mathbb{C}$ contains at least four vertices, two of which are in $\mathbb{B}$ and two of which are in $\mathbb{F}$, and $\mathbb{C}$ contains no chord from $\mathbb{B}$ to $\mathbb{F}$, it follows that there is a chordless cycle in $\mathbb{G}$ of at least four vertices. This contradicts the fact that $\mathbb{G}$ is chordal. Hence, it cannot be the case that $\mathcal{V} - \{u_1, u_2, \ldots, u_m\} \neq \emptyset$. So, $\mathbb{B}$ contains all the vertices in $\mathcal{V}$, which implies $\mathbb{G}$ has a bilateration ordering. Recall that vertices u and v of $\mathbb{B}$, which are the first two vertices of the bilateration ordering on all the vertices of $\mathbb{B}$, may be any two vertices of $\mathbb{G}$. As shown above, $\mathbb{B}$ must contain all the vertices of $\mathbb{G}$. Hence, it must be the case that for all edges (u, v) in $\mathbb{G}$, there exists a bilateration ordering of $\mathbb{G}$ that begins with vertices u and v. □

Proof of Lemma 4. If w is a proxy anchor, then the lemma holds trivially, so suppose w is not a proxy anchor. It is also easy to show that the lemma holds when u and v are both proxy anchors, so suppose at least one of u or v is not a proxy anchor. Let the ordering of the first sweep be $x_1, \ldots, x_n$, which we denote by $[x]$. Let $w = x_k$. Without loss of generality, suppose u precedes v in $[x]$, i.e., $u = x_i$ and $v = x_j$ for some i, j where $i < j$. It is easy to see that if w precedes v in the ordering $[x]$, then it cannot be the case that v is in the domain of any subassignment in $\mathcal{S}(w, 1)$. Therefore, it must be the case that $k \geq j$. We will prove the lemma by induction on $k - j$. First, consider the case where $k = j$. In this case, $w = v$. From (13), it is clear that $u \in \mathcal{D}(\beta)$ for all $\beta \in \mathcal{S}(v, 1)$, and $\|\beta(u) - \beta(v)\| = d_{uv}$. Now suppose $k - j = 1$. Let $\mathcal{M}(w) = \mathcal{N}(w) \cap \{x_1, \ldots, x_{k-1}\}$. If $x_k = w$ is not adjacent to $x_j = v$, then it is easy to see that v cannot be in the domain of any subassignment in $\mathcal{S}(w, 1)$. Hence, suppose w is adjacent to v, which implies $v \in \mathcal{M}(w)$. Let the elements of $\mathcal{M}(w)$ be denoted $u_1, \ldots, u_m$, and without loss of generality, let $v = u_1$. Let β be any subassignment of $\mathcal{S}(w, 1)$. From (15), we have that $\beta \in \mathcal{M}(\alpha_1, \ldots, \alpha_m, w, u_1, \ldots, u_m)$ for some collection of pairwise consistent subassignments $\alpha_1, \ldots, \alpha_m$ where $\alpha_i \in \mathcal{S}(u_i, 1)$, $i \in \{1, \ldots, m\}$. This implies $\mathcal{D}(\alpha_1) \subset \mathcal{D}(\beta)$. As we have just shown, $u \in \mathcal{D}(\delta)$ and $\|\delta(u) - \delta(v)\| = d_{uv}$ for all $\delta \in \mathcal{S}(v, 1)$. Hence, $u, v \in \mathcal{D}(\alpha_1)$ and $\|\alpha_1(u) - \alpha_1(v)\| = d_{uv}$. From (13), we have that the domain of each subassignment in $\mathcal{M}(\alpha_1, \ldots, \alpha_m, w, u_1, \ldots, u_m)$ contains $\mathcal{D}(\alpha_1)$ and must be identical to α_1 when restricted to the domain of α_1. Therefore, $u, v \in \mathcal{D}(\beta)$ and

$\beta(u) = \alpha_1(u)$, $\beta(v) = \alpha_1(v)$. Clearly, this implies $\|\beta(u) - \beta(v)\| = d_{uv}$. Hence, for all $\gamma \in \mathcal{S}(w, 1)$ we have that $u, v \in \mathcal{D}(\gamma)$ and $\|\gamma(u) - \gamma(v)\| = d_{uv}$.

Suppose the lemma holds for all $w = x_k$ where $k - j \leq L$ for some L. Now consider w, where $w = x_k$, where $k - j = L + 1$. Again, let $u_1, \ldots, u_m$ denote the elements of $\mathcal{M}(w)$. Let β be any subassignment of $\mathcal{S}(w, 1)$ where $u, v \in \mathcal{D}(\beta)$. From (15), we have that $\beta \in \mathcal{M}(\alpha_1, \ldots, \alpha_m, w, u_1, \ldots, u_m)$ for some collection of pairwise consistent subassignments $\alpha_1, \ldots, \alpha_m$ where $\alpha_i \in \mathcal{S}(u_i, 1)$, $i \in \{1, \ldots, m\}$. By definition, the domain of β is equal to the union of $\{w\}$ and the union of the domains of α_z, $z \in \{1, \ldots, m\}$. Hence, it must be the case that $v \in \mathcal{D}(\alpha_z)$ for some $z \in \{1, \ldots, m\}$. As noted previously, u is in the domain of all subassignments in $\mathcal{S}(v, 1)$. It is straightforward to show then that $u, v \in \mathcal{D}(\alpha_z)$. By definition of $\mathcal{M}$ in (13), we have that β must equal α_z when restricted to the domain of α_z. This implies $\beta(u) = \alpha_z(u)$ and $\beta(v) = \alpha_z(v)$. By the inductive hypothesis, we have that $\|\alpha_z(u) - \alpha_z(v)\| = d_{uv}$, which implies $\|\beta(u) - \beta(v)\| = d_{uv}$. The lemma follows by induction. $\square$

Proof of Lemma 5. It is easy to show that the lemma holds when u and v are both proxy anchors, so suppose at least of of u or v is not a proxy anchor. Let $y_1, \ldots, y_n$ denote the ordering chosen for the second sweep, and for any vertex y_i, let $\mathcal{M}(y_i) = \mathcal{N}(y_i) \cap \{y_1, \ldots, y_{i-1}\}$. Suppose $w = y_i$. We will prove the lemma by induction on $i \in \{1, \ldots, n\}$. The lemma is trivially true if w is a proxy anchor, i.e., $w = y_i$, $i \in \{1, 2, 3\}$. Now suppose the lemma holds for all $w = y_j$, where $j < i$ for some $i \in \{4, \ldots, n\}$. Consider $w = y_i$. If $\mathcal{M}(w) = \emptyset$, then $\mathcal{S}(w, 2) = \mathcal{S}(w, 1)$, in which case Lemma 5 follows from Lemma 4. So suppose $\mathcal{M}(w) \neq \emptyset$. Let $u_1, \ldots, u_m$ denote the elements of $\mathcal{M}(w)$. Let β be any subassignment of $\mathcal{S}(w, 2)$ such that $u, v \in \mathcal{D}(\beta)$. From (17), we have that $\beta = u_{m+1}(\alpha_0, \alpha_1, \ldots, \alpha_m)$ for some collection of pairwise consistent subassignments $\alpha_0 \in \mathcal{S}(w, 1)$ and $\alpha_j \in \mathcal{S}(u_j, 2)$, $j \in \{1, \ldots, m\}$. Without loss of generality, suppose that u preceded v in the ordering chosen for the first sweep. By definition of u_{m+1}, the domain of β is the union of the domains of α_j, $j \in \{0, 1, \ldots, m\}$. Hence, since $u, v \in \mathcal{D}(\beta)$, it follows that $v \in \mathcal{D}(\alpha_z)$ for some $z \in \{0, 1, \ldots, m\}$. It is straightforward to show from (17) that u must also be in $\mathcal{D}(\alpha_z)$, so $u, v \in \mathcal{D}(\alpha_z)$. From Lemma 4 and the inductive hypothesis, it follows that $\|\alpha_z(u) - \alpha_z(v)\| = d_{uv}$. Since α_j, $j \in \{0, 1, \ldots, m\}$, are pairwise consistent, it follows from the definition of u_{m+1} that $\beta(u) = \alpha_z(u)$ and $\beta(v) = \alpha_z(v)$, so $\|\beta(u) - \beta(v)\| = d_{uv}$. $\square$

Proof of Lemma 6. We first show that the lemma holds for the first sweep, i.e., for each $v \in \mathcal{V}$, there is a $\beta \in \mathcal{S}(v, 1)$ which is the restriction of $\bar{\alpha}$ to $\mathcal{D}(\beta)$, and $v \in \mathcal{D}(\beta)$. Let $v_1, \ldots, v_n$ be the ordering used to compute $\mathcal{S}(v, 1)$, $v \in \mathcal{V}$. Let v be any vertex of $\mathbb{G}$. If v is a proxy anchor, i.e., $v = v_i$ for some $i \in \{1, 2, 3\}$, then from (14), we have that $\mathcal{S}(v, 1) = \{\alpha\}$, where $\mathcal{D}(\alpha) = \{v\}$ and $\alpha(v) = \pi(v)$. Hence, the lemma holds for v_i, $i \in \{1, 2, 3\}$. Now suppose $v = v_i$, $i > 3$. From (13) and (15), we have that $v_i \in \mathcal{D}(\beta)$ for all $\beta \in \mathcal{S}(v, 1)$, so it just remains to show that there is a $\beta \in \mathcal{S}(v, 1)$ which is the restriction of $\bar{\alpha}$ to $\mathcal{D}(\beta)$. We show this by induction on v_i, $i \in \{1, 2, \ldots, n\}$. We have already shown the lemma to be true for $\mathcal{S}(v_i, 1)$, $i \in \{1, 2, 3\}$. Now suppose the lemma holds for all $\mathcal{S}(v_j, 1)$, where $j < i$ for some $i \in \{4, \ldots, n\}$. For each v_j, $j < i$, let $\bar{\beta}_{v_j}$ denote the subassignment in $\mathcal{S}(v_j, 1)$ which is the restriction of $\bar{\alpha}$ to $\mathcal{D}(\bar{\beta}_{v_j})$. Now consider v_i. Let the elements of $\mathcal{N}(v_i) \cap \{v_1, \ldots, v_{i-1}\}$ be denoted $u_1, \ldots, u_m$. Clearly, $\bar{\beta}_{u_j}$, $j \in \{1, \ldots, m\}$, are pairwise consistent, i.e., $\bar{\beta}_{u_j} \sim \bar{\beta}_{u'_j}$ for all $j, j' \in \{1, \ldots, m\}$. Consider $\mathcal{M}(\bar{\beta}_{u_1}, \ldots, \bar{\beta}_{u_m}, v_i, u_1, \ldots, u_m)$. Clearly, $\pi(v_i) \in \bigcap_{j \in \{1, \ldots, m\}} \mathcal{C}(\bar{\beta}_{u_j}(u_j), d_{v_i u_j})$. From (13), it follows that there is $\beta \in \mathcal{S}(v_i, 1)$ such that $\beta = \bar{\beta}_{u_j}$ when restricted to the domain of $\bar{\beta}_{u_j}$ for each $j \in \{1, \ldots, m\}$, and $\beta(v_i) = \pi(v_i)$. Since each $\bar{\beta}_{u_j}$ is a restriction of $\bar{\alpha}$, it follows that β must then be a restriction of $\bar{\alpha}$ as well. By induction then, we have that the lemma is true for all $\mathcal{S}(v, 1)$, $v \in \mathcal{V}$.

We have just shown that the lemma holds for the first sweep. Now we will show the lemma holds for all sweeps by induction, so suppose the lemma holds for the kth sweep, where $k \geq 1$. Let $u_1, \ldots, u_n$ be the ordering chosen for the $(k+1)$th sweep. Clearly, $\mathcal{S}(u_i, k+1) = \mathcal{S}(u_i, k)$ for $i \in \{1, 2, 3\}$, and since the lemma holds for the kth sweep, we have that the lemma holds for $\mathcal{S}(u_i, k+1)$, $i \in \{1, 2, 3\}$. Suppose the lemma holds for all $\mathcal{S}(u_j, k+1)$, where $j < i$ for some $i \in \{4, \ldots, n\}$. For each u_j, $j < i$, let $\bar{\beta}_{u_j}$ denote the subassignment in $\mathcal{S}(u_j, k+1)$ which is a restriction of $\bar{\alpha}$ with u_j in its domain. Consider u_i. Clearly, if $\mathcal{N}(u_i) \cap \{u_1, \ldots, u_{i-1}\}$ is the empty set, then $\mathcal{S}(u_i, k+1) = \mathcal{S}(u_i, k)$, in which case the lemma holds for $\mathcal{S}(u_i, k+1)$ since the lemma is true for the kth sweep. So suppose $\mathcal{N}(u_i) \cap \{u_1, \ldots, u_{i-1}\} \neq \emptyset$, and let $w_1, \ldots, w_m$ denote its elements. By the inductive hypothesis, we have subassignments $\bar{\beta}_{u_i} \in \mathcal{S}(u_i, k)$ and $\bar{\beta}_{w_j} \in \mathcal{S}(w_j, k+1)$, $j \in \{1, \ldots, m\}$, where each subassignment is a restriction of $\bar{\alpha}$ and $u_i \in \mathcal{D}(\bar{\beta}_{u_i})$, $w_j \in \mathcal{D}(\bar{\beta}_{w_j})$, $j \in \{1, \ldots, m\}$. From (17), it is easy to see that $u_{m+1}(\bar{\beta}_{u_i}, \bar{\beta}_{w_1}, \ldots, \bar{\beta}_{w_m})$ is in $\mathcal{S}(u_i, k+1)$ and is also a restriction of $\bar{\alpha}$ with u_i in its domain. By induction, the lemma holds for the $(k+1)$th sweep. □

Proof of Lemma 7. Let $\mathbb{H}'_i$ denote the graph induced in $\mathbb{G}$ by the vertices of $\mathbb{H}_i$ and the vertices of $\mathcal{A}$. Suppose $\mathbb{H}'_i$ is not globally rigid. Consider the subnetwork of $\mathbb{N}$ containing just the sensors corresponding to vertices in $\mathbb{H}'_i$, and denote the subnetwork by $\mathbb{N}_i$. Clearly, the point formation modeling $\mathbb{N}_i$ is $(\mathbb{H}'_i, p')$, where p' contains the positions of those sensors of $\mathbb{N}_i$. Since the multipoint of $\mathbb{N}$ is generic, it follows that the multipoint of $\mathbb{N}_i$, i.e., p', must also be generic. Hence, that $\mathbb{H}'_i$ is not globally rigid implies $\mathbb{N}_i$ cannot be localizable. In other words, there exists multipoint q' such that the point formations $(\mathbb{H}'_i, q')$ and $(\mathbb{H}'_i, p')$ have the same edge lengths but are not congruent. Furthermore, it is easy to see that by applying a Euclidean transformation to the points of $(\mathbb{H}'_i, q')$, we can obtain a point formation $(\mathbb{H}'_i, q'')$, which is congruent to $(\mathbb{H}'_i, q')$, and such that the points in q'' corresponding to the vertices in $\mathcal{A}$ are identical to the points in p' corresponding to the vertices in $\mathcal{A}$. Hence, $(\mathbb{H}'_i, q'')$ has the same edge lengths as $(\mathbb{H}'_i, p')$, and the points corresponding to vertices in $\mathcal{A}$ are the same in both $(\mathbb{H}'_i, q'')$ and $(\mathbb{H}'_i, p')$, but $(\mathbb{H}'_i, q'')$ and $(\mathbb{H}'_i, p')$ are not congruent. Let $(\mathbb{G}, p)$ be the point formation modeling $\mathbb{N}$. Since $\mathbb{N}$ is localizable, it follows that $(\mathbb{G}, p)$ is globally rigid. Consider the point formation $(\mathbb{G}, p'')$ defined as follows. The point in $(\mathbb{G}, p'')$ corresponding to a vertex j not in $\mathbb{H}_i$ is the same as the point corresponding to vertex j in $(\mathbb{G}, p)$, and the point in $(\mathbb{G}, p'')$ corresponding to a vertex j in $\mathbb{H}_i$ is the same as the point corresponding to vertex j in $(\mathbb{H}'_i, q'')$. It is easy to see that $(\mathbb{G}, p'')$ has the same edge lengths as $(\mathbb{G}, p)$ but $(\mathbb{G}, p'')$ and $(\mathbb{G}, p)$ are not congruent. This contradicts the fact that $(\mathbb{G}, p)$ is globally rigid, and therefore $\mathbb{H}'_i$ must be globally rigid. □

Proof of Lemma 8. In the following, we will show that the lemma holds for the case where $c = 1$; i.e., the graph induced in $\mathbb{G}$ by vertices not in $\mathcal{A}$ is connected. The case for $c > 1$ follows easily. Let $\mathbb{H}$ denote the graph induced in $\mathbb{G}$ by vertices not in $\mathcal{A}$. Let the ordering for the second sweep be $[x] = x_1, \ldots, x_n$, and suppose $[x]$ is a complete ordering of $\mathbb{H}$ with respect to v of $\mathbb{H}$ and $\mathcal{A}$. This implies $v = x_n$, and $[x]$ is also a complete ordering of $\mathbb{G}$ with respect to $\mathcal{A}$ and x_n.

Let x_i be any vertex such that there is a path from x_i to x_n in $\mathbb{G}$ which is a subsequence of $[x]$ beginning with a vertex which is not a proxy anchor. In other words, there exists $i < i_1 < i_2 < \cdots < i_p < n$ such that $i_1 > 3$ and $(x_i, x_{i_1}), (x_{i_1}, x_{i_2}),$

$\dots, (x_{i_p}, x_n) \in \mathcal{E}$. We will show by induction that $x_i \in \mathcal{D}(\beta)$ for all $\beta \in \mathcal{S}(x_n, 2)$. For notational convenience, let $i_{p+1} = n$. For any x_j, let $\mathcal{M}(x_j) = \mathcal{N}(x_j) \cap \{x_1, \dots, x_{j-1}\}$. Clearly, $x_i \in \mathcal{M}(x_{i_1})$. From (17), it follows that $x_i \in \mathcal{D}(\beta)$ for all $\beta \in \mathcal{S}(x_{i_1}, 2)$. Now suppose $x_i \in \mathcal{D}(\beta)$ for all $\beta \in \mathcal{S}(x_{i_j}, 2)$, where $j \leq I$ for some $I < p + 1$, and consider $\mathcal{S}(x_{i_{j+1}}, 2)$. Since $x_{i_j} \in \mathcal{M}(x_{i_{j+1}})$, it follows from (17) that for all $\beta \in \mathcal{S}(x_{i_{j+1}}, 2)$, it must be the case that $\mathcal{D}(\beta') \subseteq \mathcal{D}(\beta)$, where β' is some subassignment of $\mathcal{S}(x_{i_j}, 2)$. But since $x_i \in \mathcal{D}(\beta')$ for all $\beta' \in \mathcal{S}(x_{i_j}, 2)$, it follows that $x_i \in \mathcal{D}(\beta)$ for all $\beta \in \mathcal{S}(x_{i_{j+1}}, 2)$. By induction then, we have that $x_i \in \mathcal{D}(\beta)$ for all $\beta \in \mathcal{S}(x_n, 2)$.

Let $\mathcal{N}_0(x_n), \dots, \mathcal{N}_h(x_n)$ be the vertex partition of $\mathbb{H}$ used to construct the complete ordering $[x]$. Consider any x_i, $i > 3$, and suppose $x_i \in \mathcal{N}_j(x_n)$. Now we show that there is a path from x_i to x_n in $\mathbb{G}$ which is a subsequence of $[x]$; i.e., there exists $i < i_1 < i_2 < \cdots < i_p < n$ such that $(x_i, x_{i_1}), (x_{i_1}, x_{i_2}), \dots, (x_{i_p}, x_n) \in \mathcal{E}$. Since $x_i \in \mathcal{N}_j(x_n)$, it must be true that x_i is adjacent to some vertex in $\mathcal{N}_{j'}(x_n)$, where $j' < j$. But since all the vertices in sets $\mathcal{N}_{j'}$, $j' < j$, are assigned larger indices than vertices in $\mathcal{N}_j$, it follows that x_i must be adjacent to some vertex x_{i_1} where $i < i_1$. If $x_{i_1} \in \mathcal{N}_0$, then it must be the case that $x_n = x_{i_1}$. Otherwise, $x_{i_1} \in \mathcal{N}_b$ where $b > 0$, and so x_{i_1} must be adjacent to some vertex x_{i_2} in $\mathcal{N}_a$ where $a < b$. By construction of a complete ordering, we have that $i_1 < i_2$ since $x_{i_2} \in \mathcal{N}_a$ and $a < b$. Hence, there must exist a sequence of vertices $x_{i_1}, \dots, x_{i_p} = x_n$ such that x_i is adjacent to x_{i_1}, each x_{i_j} is adjacent to $x_{i_{j+1}}$, and $i < i_1 < i_2 < \cdots < i_p = n$.

From the above, we can conclude that each x_i, $i > 3$, must be in $\mathcal{D}(\beta)$ for all $\beta \in \mathcal{S}(x_n, 2)$. Now consider the proxy anchors, i.e., x_i, $i \in \{1, 2, 3\}$. Suppose some x_i, $i \in \{1, 2, 3\}$, is not adjacent to any x_j, $j > 3$. This implies x_i has degree two in $\mathbb{G}$, and therefore $\mathbb{G}$ cannot be globally rigid in $\mathbb{R}^2$, and $\mathbb{N}$ is not localizable. This is clearly a contradiction. Hence, each x_i, $i \in \{1, 2, 3\}$, must be adjacent to some x_j, $j > 3$, which implies there exist indices $i_1, \dots, i_p$, where $i < i_1 < i_2 < \cdots < i_p < n$, $i_1 > 3$, and $(x_i, x_{i_1}), (x_{i_1}, x_{i_2}), \dots, (x_{i_p}, x_n) \in \mathcal{E}$. Hence, each x_i, $i \in \{1, 2, 3\}$, must also be in $\mathcal{D}(\beta)$ for all $\beta \in \mathcal{S}(x_n, 2)$, and it follows that $\mathcal{D}(\beta) = \mathcal{V}$ for all $\beta \in \mathcal{S}(x_n, 2)$. □

Proof of Lemma 9. First, suppose $c = 1$ so the graph $\mathbb{H}$ induced in $\mathbb{G}$ by vertices which do not correspond to the proxy anchors is connected. Suppose the ordering used to compute the second sweep is a complete ordering of $\mathbb{G}$ with respect to $\mathcal{A}$ and vertex u of $\mathbb{H}$. From Lemma 6, we have that $\mathcal{S}(u, 2)$ is not empty. From Lemma 8, we have that for each $\alpha \in \mathcal{S}(u, 2)$, the domain of α is equal to $\mathcal{V}$. From Lemma 5 we have that $\|\alpha(u) - \alpha(v)\| = d_{uv}$ for all $(u, v) \in \mathcal{E}$. Clearly, $\alpha(a) = \pi(a)$ for all proxy anchors a. Hence, α is a consistent assignment of $\mathbb{N}$, where $\alpha(a) = \bar{\alpha}(a)$ for all proxy anchors a. But as noted previously, there can be at most one such assignment, which implies that α must equal $\bar{\alpha}$. Now we consider the case for $c > 1$. It follows from Lemma 7 that each subnetwork $\mathbb{N}_i$ containing sensors corresponding to vertices in $\mathcal{A}$ and $\mathbb{H}_i$ is itself localizable. The argument for the case $c = 1$ can be applied to each $\mathbb{N}_i$ to show that each subassignment in $\mathcal{S}(u_i, 2)$ must be the restriction of $\bar{\alpha}$ to the vertices corresponding to the proxy anchors and the sensors in $\mathbb{N}_i$. □

Proof of Theorem 2. We prove the lemma for the case where $c = 1$; i.e., the graph $\mathbb{G}(\mathcal{V} - \mathcal{W})$ is connected. The lemma for the case where $c > 1$ is a direct consequence.

Since $v_1, \dots, v_n$ is the ordering chosen for the first sweep, without loss of generality we can suppose v_1, v_2, v_3 are the proxy anchors and let $\mathcal{W} = \{v_1, v_2, v_3\}$. Since $\mathbb{G}$ is partially acyclic and $c = 1$, it follows that $\mathbb{G}(\mathcal{V} - \mathcal{W})$ is acyclic and connected. Hence, for each $v \in \mathcal{V} - \mathcal{W}$, we can define $l(v)$, where $l(v)$ is the length of the path from v to v_n in $\mathbb{G}(\mathcal{V} - \mathcal{W})$. Clearly $l(v_n) = 0$. Let $L = \max_{v \in \mathcal{V} - \mathcal{W}} l(v)$. For each v_i, let $\mathcal{M}(v_i) = \mathcal{N}(v_i) \cap \{v_1, \dots, v_{i-1}\}$. Let $[v]$ denote the ordering $v_1, \dots, v_n$. Let p be any point in $\mathcal{S}(v_n, 1)$. In the following we will assign point $p(v)$ to each sensor $v \in \mathcal{V} - \mathcal{W}$ so that all known intersensor distances are satisfied and the point assigned to v is p, i.e., $p(v_n) = p$. We do this inductively on $l(v)$ beginning with v where $l(v) = 0$. Obviously v_n is the only vertex such that $l(v_n) = 0$, and we let $p(v_n) = p$. Now we consider v where $l(v) = 1$. As noted previously, $\mathcal{M}(v_n) - \mathcal{W} = \mathcal{N}_1(v_n)$, and by definition $\mathcal{N}_1(v_n)$ is the set of vertices v where $l(v) = 1$. Let $u_1, \ldots, u_m$ denote the vertices in $\mathcal{M}(v_n) - \mathcal{W}$. From (2), there are points $p_i \in \mathcal{S}(u_i, 1)$ such that $\|p - p_i\| = d_{v_n u_i}$. For each u_i, $i \in \{1, \ldots, m\}$, let $p(u_i) = p_i$. Now suppose $p(v)$ has been defined for all vertices v where $l(v) \leq k$ for some $k < L$. Now we define $p(v)$ for each vertex $v \in \mathcal{V} - \mathcal{W}$ where $l(v) = k + 1$. Since $[v]$ is a complete ordering and $\mathbb{G}(\mathcal{V} - \mathcal{W})$ is acyclic, we have that if vertex v is such that $l(v) = k+1$, then there must exist exactly one vertex $v' \in \mathcal{V} - \mathcal{W}$ such that $(v, v') \in \mathcal{E}$, and $l(v') \leq k$. Since $l(v') = k$, it follows that $p(v')$ has already been defined. Furthermore, from (2), we have that there must exist point $p_v \in \mathcal{S}(v, 1)$ such that $\|p(v') - p_v\| = d_{vv'}$. Let $p(v) = p_v$.

Let u, v be any two vertices in $\mathcal{V} - \mathcal{W}$ which are adjacent in $\mathbb{G}$. Since $\mathbb{G}(\mathcal{V} - \mathcal{W})$ is acyclic, it follows that either $l(u) = l(v) + 1$ or $l(v) = l(u) + 1$, which implies $\|p(u) - p(v)\| = d_{uv}$. Now let w be any vertex of $\mathcal{W}$, and let u be any vertex of $\mathcal{V} - \mathcal{W}$. Since all the sensors in $\mathcal{W}$ are proxy anchors, each $w \in \mathcal{W}$ is assigned some position $\pi(w)$ by the first sweep of the Restricted Sweeps algorithm. Again from (2), we have that for all $p_u \in \mathcal{S}(u, 1)$, it must be the case that $\|p(u) - \pi(w)\| = d_{uw}$. Hence, if we define $p(w) = \pi(w)$ for all $w \in \mathcal{W}$, and assigned position $p(v)$ to each sensor $v \in \mathcal{V}$, then all known intersensor distances must be satisfied. We have just shown that for each $p \in \mathcal{S}(v_n, 1)$, there correspond points $p(v)$, $v \in \mathcal{V}$, such that $p(v_n) = p$ and $p(w) = \pi(w)$ for all $w \in \mathcal{W}$, and all known intersensor distances are satisfied. Since $\mathbb{N}$ is localizable, we have that for all assignments of points $q(v)$ to sensors $v \in \mathcal{V} - \mathcal{W}$ such that all known intersensor distances are satisfied, assuming each sensor $w \in \mathcal{W}$ is positioned at $\pi(w)$, it must be the case that $q(v) = p(v)$ for all $v \in \mathcal{V} - \mathcal{W}$. This implies $\mathcal{S}(v_n, 1)$ can contain only one element. Let $\pi(v_n)$ denote the point in $\mathcal{S}(v_n, 1)$. Clearly, $\pi(v_n)$ is the position of sensor v_n relative to the positions assigned to the proxy anchors. Let $p(v)$, $v \in \mathcal{V}$, be as defined above. Now we show that $\mathcal{S}(v, 2)$, $v \in \mathcal{V}$, must all be singletons. This is trivially true for $v \in \mathcal{W} \cup \{v_n\}$, so consider $v \notin \mathcal{W} \cup \{v_n\}$. Since $\mathbb{G}(\mathcal{V} - \mathcal{W})$ is acyclic, it follows that v is adjacent to exactly one v' in $\mathbb{G}(\mathcal{V} - \mathcal{W})$ such that $l(v') < l(v)$. By definition, $\|p(v) - p(v')\| = d_{vv'}$, and the only criterion used for choosing $p(v)$ from $\mathcal{S}(v, 1)$ was that $\|p(v) - p(v')\| = d_{vv'}$. This implies that if there is $q \in \mathcal{S}(v, 1)$ where $q \neq p(v)$ and $\|q - p(v')\| = d_{vv'}$, then there exists an assignment of points $q(x)$, $x \in \mathcal{V}$, such that $q(v) = q$, $q(w) = \pi(w)$ for all $w \in \mathcal{W}$, and all known intersensor distances are satisfied. But this clearly contradicts the assumption that $\mathbb{N}$ is localizable. Hence, there can exist only one point $p \in \mathcal{S}(v, 1)$, namely $p(v)$, such that $\|p - p(v')\| = d_{vv'}$. From (3) and the ordering specified for the second sweep, it follows that $\mathcal{S}(v, 2)$ must be a singleton consisting of only $p(v)$. Hence, $\mathcal{S}(v, 2)$ must be a singleton for all $v \in \mathcal{V}$. □

Proof of Lemma 13. Without loss of generality, we assume that v_1, v_2, and v_3 are anchors. Note that the coordinates computed accordingly for the remaining sensors can be transformed into their real locations by Euclidean transformations since there are three anchors in the network. Now consider the first sweep with the ordering $v_1, v_2, \ldots, v_n$ in $\bar{\mathbb{G}}^2$. Since v_4 is adjacent to both v_2 and v_3, we have that $\mathcal{S}(v_4, 1)$ contains two elements. Since v_5 is adjacent to both v_4 and v_3, we have that $\mathcal{S}(v_5, 1)$ contains four elements. Similarly, $\mathcal{S}(v_i, 1)$, $4 \leq i \leq n$ contains finite elements by using the edges (v_i, v_{i-1}) and (v_i, v_{i-2}) in $\bar{\mathbb{G}}^2$. Then consider the second sweep with the ordering $v_1, v_2, v_3, v_n, v_{n-1}, \ldots, v_4$. From Lemma 2.1 in [14] we know that $\bar{\mathbb{G}}^2$ is generically globally rigid, which implies that generically there is one element in

$\mathcal{S}(v_n, 1)$ which satisfies simultaneously $||p_{v_n} - p_{v_1}|| = d_{v_n v_1}$ and $||p_{v_n} - p_{v_2}|| = d_{v_n v_2}$. Hence, $\mathcal{S}(v_n, 2)$ contains exactly one element. Using the same reasoning, we know that $\mathcal{S}(v_{n-1}, 2)$ contains one element by using the edges (v_{n-1}, v_n) and (v_{n-1}, v_1) in $\bar{\mathbb{G}}^2$. Similarly, we know that $\mathcal{S}(v_i, 2)$, $4 \leq i \leq n-2$ by using the edges (v_i, v_{n+1}) and (v_i, v_{i+2}) in $\bar{\mathbb{G}}^2$. □

REFERENCES

[1] B. D. O. Anderson, P. N. Belhumeur, T. Eren, D. K. Goldenberg, A. S. Morse, W. Whiteley, and Y. R. Yang, *Graphical properties of easily localizable sensor networks*, Wireless Networks, 2007 (electronic).

[2] J. Aspnes, T. Eren, D. K. Goldenberg, A. S. Morse, W. Whiteley, Y. R. Yang, B. D. O. Anderson, and P. N. Belhumeur, *A theory of network localization*, IEEE Trans. Mobile Computing, 5 (2006), pp. 1663–1678.

[3] P. Biswas, T.-C. Lian, T.-C. Wang, and Y. Ye, *Semidefinite programming based algorithms for sensor network localization*, ACM Trans. Sensor Networks, 2 (2006), pp. 188–220.

[4] R. Connelly, *Generic global rigidity*, Discrete Comput. Geom., 33 (2005), pp. 549–563.

[5] T. Eren, D. Goldenberg, W. Whiteley, Y. R. Yang, A. S. Morse, B. D. O. Anderson, and P. N. Belhumeur, *Rigidity, computation, and randomization in network localization*, in Proceedings of the 23rd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM), New Haven, CT, 2004, pp. 2673–2684.

[6] J. Fang, D. Duncan, and A. S. Morse, *Sequential localization with inaccurate measurements*, to appear in Localization Algorithms and Strategies for Wireless Sensor Networks: Monitoring and Surveillance Techniques for Target Tracking.

[7] D. K. Goldenberg, P. Bihler, M. Cao, J. Fang, B. D. O. Anderson, A. S. Morse, and Y. R. Yang, *Localization in sparse networks using sweeps*, in Proceedings of Mobicom, 12th Annual International Conference on Mobile Computing and Networking, ACM, New York, 2006, pp. 110–121.

[8] J. E. Graver, B. Servatius, and H. Servatius, *Combinatorial Rigidity*, Grad. Stud. Math. 2, AMS, Providence, RI, 1993.

[9] B. Hendrickson, *Conditions for unique graph realizations*, SIAM J. Comput., 21 (1992), pp. 65–84.

[10] B. Jackson and T. Jordán, *Connected rigidity matroids and unique realizations of graphs*, J. Combin. Theory Ser. B, 94 (2005), pp. 1–29.

[11] D. J. Jacobs and B. Hendrickson, *An algorithm for two-dimensional rigidity percolation: The pebble game*, J. Comput. Phys., 137 (1997), pp. 346–365.

[12] D. Moore, J. Leonard, D. Rus, and S. Teller, *Robust distributed network localization with noisy range measurements*, in Proceedings of the 2nd ACM SenSys, Baltimore, MD, 2004, pp. 50–61.

[13] N. B. Priyantha, H. Balakrishnan, E. Demaine, and S. Teller, *Poster abstract: Anchor-free distributed localization in sensor networks*, in SenSys '03: Proceedings of the 1st International Conference on Embedded Networked Sensor Systems, ACM, New York, 2003, pp. 340–341.

[14] A. Savvides, H. Park, and M. Srivastava, *The n-hop multilateration primitive for node localization problems*, ACM Mobile Networks Appl., 8 (2003), pp. 443–451.

[15] Y. Shang and W. Ruml, *Improved MDS-based localization*, in Proceedings of the 23rd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM), New Haven, CT, 2004, pp. 2640–2651.

[16] Y. Shang, W. Ruml, Y. Zhang, and M. P. J. Fromherz, *Localization from mere connectivity*, in MobiHoc '03: Proceedings of the 4th ACM International Symposium on Mobile Ad Hoc Networking and Computing, ACM, New York, 2003, pp. 201–212.

[17] T. Tay and W. Whiteley, *Generating isostatic frameworks*, Structural Topology, 306 (1988), pp. 115–139.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 1, pp. 351–372

© 2009 Society for Industrial and Applied Mathematics

NONUNIFORM COVERAGE AND CARTOGRAMS*

FRANCOIS LEKIEN† AND NAOMI EHRICH LEONARD‡

Abstract. In this paper, we investigate nonuniform coverage of a planar region by a network of autonomous, mobile agents. We derive centralized nonuniform coverage control laws from uniform coverage algorithms using cartograms, transformations that map nonuniform metrics to a near Euclidean metric. We also investigate time-varying coverage metrics and the design of control algorithms to cover regions with slowly varying, nonuniform metrics. Our results are applicable to the design of mobile sensor networks, notably when the coverage metric varies as data is collected such as in the case of an information metric. The results apply also to the study of animal groups foraging for food that is nonuniformly distributed and possibly changing.

Key words. cartogram, optimal coverage, mobile sensor networks, adaptive sampling

AMS subject classifications. 49K99, 65J10, 35G10

DOI. 10.1137/070681120

1. Introduction.

Sensor networks in space, in the air, on land, and in the ocean provide the opportunity for unprecedented observational capability. An important problem in this context is to determine how best to distribute sensors over a given area in which the observational field is distributed so that the likelihood of detecting an event of interest is maximized. If the probability distribution of the event is uniform over the area, then the optimal solution is uniform coverage, i.e., uniform distribution of sensors. On the other hand, if this probability distribution is nonuniform, then the sensors should be more (less) densely distributed in subregions with higher (lower) event probability. Further, if the probability distribution changes with time, then the nonuniform distribution should likewise change with time.

A related coverage problem derives from the classic objective analysis (OA) mapping error in problems of sampling (possibly time-varying) scalar fields, e.g., temperature in the ocean [3]. OA is linear statistical estimation based on specified field statistics, and the mapping error provides a measure of statistical uncertainty of the model as a function of where and when the data is taken. Since reduced uncertainty, equivalent to increased entropic information, implies better measurement coverage, OA mapping error can be used as a coverage metric [3, 15]. If the a priori error correlation between any two points in the plane is homogeneous and isotropic, then uniform coverage will be optimal initially. However, the optimal coverage solution will not be static (unless the scalar field of interest changes very quickly), since, once a particular location has been sampled, it should not be sampled again until the measurement value has decayed sufficiently.

Robotic vehicles carrying sensors in space, in the air, on land, and in the ocean make possible mobile sensor networks that can adapt to changing coverage requirements. Given a coverage metric that is independent of time and of the history of

*Received by the editors January 26, 2007; accepted for publication (in revised form) August 10, 2008; published electronically February 11, 2009. This research was supported by the Jules Reyers Foundation and ONR grants N00014-02-1-0826 and N00014-04-1-0534.
http://www.siam.org/journals/sicon/48-1/68112.html

†École Polytechnique CP 165/11, Université Libre de Bruxelles, B-1050 Brussels, Belgium (lekien@ulb.ac.be).

‡Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ 08544 (naomi@princeton.edu).

samples taken, the goal is to design coordinated control dynamics for the vehicles that yield convergence to the maximum coverage configuration from arbitrary initial conditions. A second objective is to extend these coordinated control dynamics to the case in which the coverage metric definition changes in time or, as in the case of the OA error metric, as a function of past measurement locations and times.

Coverage problems have a compelling analogy in possible models of social foraging by animal groups. Backed by observations of animal behavior across a number of species, biologists model distribution of animals over patchy resource environments according to a measure of patch suitability that depends on factors such as resource richness or conditions for survival [7, 2, 20]. Suitability varies with animal density; a typical assumption is that suitability decreases with increasing animal population. For example, suitability declines when more animals converge on a given patch since resources (e.g., prey) may be limited, and thus average consumption rates go down with more hungry consumers. Since animals prefer patches with higher suitability, nonuniformity in the distribution of resources (i.e., the suitability) reflects nonuniformity in the distribution of the animal group. As in the case of the OA mapping error metric that changes as samples are taken, suitability decreases in time as animals consume (and animals will abandon patches where suitability has declined). By this analogy, coverage studies of changing, nonuniform environments may prove useful in helping to explain how animal groups move and redistribute.

Several contributions have been made to the design of coverage algorithms for a group of dynamic agents, including [4, 18, 17, 5] and the references therein. In [4], uniform coverage algorithms are derived using Voronoi cells and gradient laws for distributed dynamical systems. Uniform constrained coverage control is addressed in [18], where the constraint is a minimum limit on node degree. Node degree refers to the number of neighbors for each agent, where a neighbor is any other agent that is sufficiently close by. Virtual potentials enable repulsion between agents to maximize coverage and attraction between agents to enforce the constraint. In [17], gradient control laws are proposed to move sensors to a configuration that maximizes expected event detection frequency. Local rules are enforced by defining a sensing radius for each agent, which also makes computations simpler. The approach is demonstrated for a nonuniform but symmetric density field with and without communication constraints. Further results for distributed coverage control are presented in [5] for a coverage metric defined in terms of the Euclidean metric with a weighting factor that allows for nonuniformity. As in [4], the methodology makes use of Voronoi cells and Lloyd descent algorithms.

In this paper we concentrate exclusively on two-dimensional planar regions, and we propose an approach to coverage control that makes use of existing algorithms designed for uniform coverage and extends these to nonuniform metrics. We are particularly interested in metrics defined in terms of non-Euclidean distance functions that effectively stretch and shrink space in lower and higher density regions of a given space. This yields optimal configurations where the resource or information is evenly distributed among the agents. Non-Euclidean distance metrics present challenges to existing techniques. For example, in the case of [4, 5], computing Voronoi cells with non-Euclidean metrics is computationally complex. For each point on a dense grid, one needs to compute the (non-Euclidean) distance to each agent and find the minimum. Computing Voronoi cells for a non-Euclidean metric is therefore much more demanding than the corresponding problem with an Euclidean metric where the polygonal boundaries of the Voronoi cells can be computed directly.

The first step in our method is to compute a nonuniform change of coordinates on the original compact set with a non-Euclidean metric that maps to a new compact set with a near Euclidean metric. Such a map is called a *cartogram.* Inspired by the work of Gastner and Newman [9], we compute the cartogram from a diffusion equation. Gastner and Newman used cartograms in several applications, including the representation of election results [11] and the optimal design of spatial distribution networks [10]. For these problems, it is sufficient to compute single cartograms, while we are interested in computing a series of cartograms for feedback control. Accordingly, we propose a method to compute cartograms that vary smoothly as a function of the density distribution.

A uniform control law can be used in the cartogram space since the metric in this space is almost Euclidean. The preimage of the control law yields convergent dynamics in the original space. We prove under certain conditions that these convergent dynamics optimize the nonuniform coverage metric. We show how to extend the approach to the case of a time-varying metric.

A limitation of our approach is the centralized computation of the cartogram. However, we note that the density function does not need to be known a priori; it can be measured or computed by the agents in real time. For example, the changing OA metric can be computed only on the fly since it is a function of where and when samples have been taken.

In section 2 we review, as an example, the uniform coverage control of [4]. We describe the nonuniform coverage problem in section 3. Cartograms are defined in section 4. Gastner and Newman's method for computing cartograms using the diffusion equation is reviewed, and our new approach to computing smooth cartograms is presented. In section 5 we describe and prove our approach to nonuniform coverage control that makes use of cartograms. Section 6 provides error estimates for the examples studied, and the case of time-varying metrics is addressed in section 7. Conclusions and future directions are given in section 8.

2. Uniform coverage.

A number of different metrics and different coordinated control strategies have been developed for uniform coverage, as described above. In this section, as a motivating and useful example, we review the uniform coverage approach and result of Cortés and Bullo, who devised a robust and efficient control scheme to optimize the configuration of a group of robotic vehicles carrying sensors [4]. They consider a group of n vehicles moving in a region $\mathcal{D}$, with a polygonal boundary $\partial\mathcal{D}$. The vehicles obey first-order dynamics:

$$\dot{\mathbf{x}}_i = \mathbf{u}_i(\mathbf{x}_1, \dots, \mathbf{x}_n), \tag{2.1}$$

where $\mathbf{x}_i$ is the position of the ith vehicle and $\mathbf{u}_i$ is the control input to the ith vehicle.

The goal is to bring the robots, from their initial positions, to a (static) configuration that maximizes coverage of the domain. To define *maximum coverage*, Cortés and Bullo consider multicenter metric functions such as

$$\Phi(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) = \max_{\mathbf{x}\in\mathcal{D}} \left\{ \min_{i=1\cdots n} d(\mathbf{x}, \mathbf{x}_i) \right\}, \tag{2.2}$$

where $d(\mathbf{x}, \mathbf{x}_i) = \|\mathbf{x} - \mathbf{x}_i\|$ is the Euclidean distance. Given the position of the n vehicles, computing the metric requires computing the distance from any point $\mathbf{x} \in \mathcal{D}$ to the closest vehicle. The metric Φ is equal to the largest of these distances. As a result, the maximum distance between any point of the domain and the closest vehicle is always smaller than or equal to Φ. Intuitively, a smaller Φ implies that the corresponding array of vehicles $\mathbf{x}_i$ achieves a better coverage of the domain $\mathcal{D}$.

Assuming that all of the vehicles have the same constant speed, Φ is proportional to the maximum time it takes for a vehicle to reach an arbitrary point of the domain. For this reason, Cortés and Bullo define optimal coverage as the minimum of the cost function Φ.

One of the main results of [4] is the development of a stable procedure to bring the vehicles into a configuration that minimizes the metric Φ. To this end, the Voronoi cell of each vehicle is computed repeatedly. The Voronoi cell for the ith vehicle is a polygonal subset of the domain $\mathcal{D}$ that contains all of the points that are closer to the ith vehicle than any other vehicle. Each vehicle is then directed to move toward the circumcenter of its Voronoi cell (see Figure 1). Once all the vehicles reach the circumcenter of their Voronoi cell, the coverage metric Φ is minimum (see the last panel of Figure 1). Cortés and Bullo show that, from any initial position where the vehicles are not exactly on top of each other, their algorithm converges toward the optimal configuration.

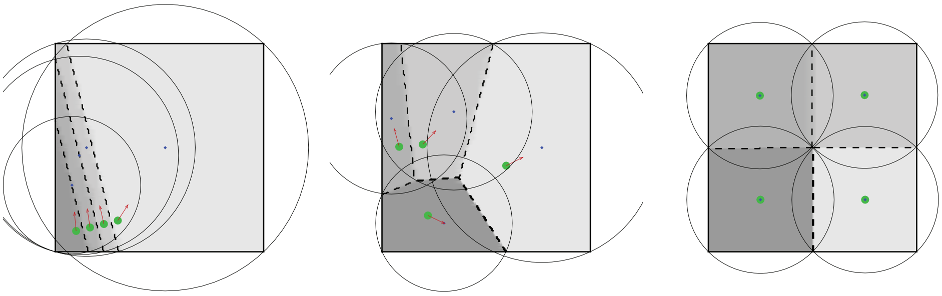


FIG. 1. *Convergence to static uniform coverage.* Thick dots: *Position of the four agents.* Shaded polygons: *Voronoi cell for each agent.* Large circles: *Circumcircle of each Voronoi cell.* Diamonds: *Centers of each circumcircle (i.e., circumcenters).* Arrows: *Velocity of the vehicles (oriented along the segment joining the agents to the circumcenter of their Voronoi cell).*

3. Nonuniform coverage. In the present paper we develop an approach that extends optimal coverage strategies to more general metrics, notably to nonuniform and time-varying metrics. Nonuniform metrics are motivated by coverage problems in environments with nonuniformly distributed information or resources. The objective is to produce distributions of agents that match the inhomogeneities in the information or resource field. We are particularly interested in metrics defined in terms of a (possibly time-varying) distance function that is non-Euclidean. A non-Euclidean distance function stretches high density areas and shrinks low density areas. With a coverage metric that depends on such a distance function, individual agents can be organized so that information or resources are equally distributed among them. In other words, each agent has a "dominance region" (e.g., Voronoi cell) and, irrespective of the size of these regions, we want the amount of information or the resource to be equal in each agent's cell.

Notice that this objective is slightly different from the approach of Cortés and Bullo. In [4], the region of dominance of an agent is flexible and might include points that can be reached more easily by other agents. In this paper, we consider the dominance region of an agent $\mathbf{x}_i$ as a defined region containing all of the points that are closer to $\mathbf{x}_i$ than any other agent in the sense of the non-Euclidean metric, i.e., that can be reached more easily by agent $\mathbf{x}_i$ than by any other agent, where ease in reaching a point depends on density. In other words, the dominance region is still a Voronoi cell, and the nonuniform density is introduced through the distance function used to compute the Voronoi cells. The nonuniform distance shrinks along paths where resources are sparse and increases along paths where resources are plentiful.

Recall that the Euclidean distance between two points is also the length of the shortest path between the two points, or

$$\|\mathbf{x}-\mathbf{x}_i\| = \min_{\mathcal{C}_{\mathbf{x}}^{\mathbf{x}_i}} \left\{ \int_{\mathcal{C}_{\mathbf{x}}^{\mathbf{x}_i}} \mathrm{dl} \right\},$$

where $\mathcal{C}_{\mathbf{x}}^{\mathbf{x}_i}$ is an arbitrary path from $\mathbf{x}$ to $\mathbf{x}_i$. If the density of information $\rho : \mathcal{D} \to \mathbb{R}_0^+$ is not uniform, then we can define a non-Euclidean distance:

$$d_\rho(\mathbf{x},\mathbf{x}_i) = \min_{\mathcal{C}_{\mathbf{x}}^{\mathbf{x}_i}} \left\{ \int_{\mathcal{C}_{\mathbf{x}}^{\mathbf{x}_i}} \sqrt{\rho}\, \mathrm{dl} \right\}.$$

The distance between two points in a low-density area is less than the Euclidean distance. The shortest path between two points might be curved in order to avoid peaks of ρ. High density implies that the distances are stretched; hence more vehicles are needed in the area. The region of dominance of each agent is still its Voronoi cell, but the nonuniform distance changes the shape of the cell. As an example see the bottom right panel of Figure 6, which shows four vehicles distributed optimally with respect to a non-Euclidean metric; in this example, the peak density is in the lower right corner of the region.

Note that we use the square root of the function ρ and not the function ρ itself to weight the distance integral. The reason for this choice is the fact that, in two dimensions, multiplying the distances in each direction by $\sqrt{\rho}$ implies a net volume (or density) change of ρ. It is also worth noting that weighting the distance integral by a negative function is not acceptable, as some distances would become negative.

In this paper we assume that the coverage metric Φ is a functional of a distance function d_ρ, which depends on the positions of the agents $\mathbf{x}_i$ and the domain $\mathcal{D}$, denoted

$$\Phi = \left(\Phi[d_\rho]\right)(\mathbf{x}_i,\ldots,\mathbf{x}_n;\mathcal{D}). \tag{3.1}$$

Clearly, one can use any metric Φ that involves only the Euclidean distance, such as the multicenter function (2.2), and make it inhomogeneous by replacing the Euclidean distance d with the weighted distance d_ρ. If ρ represents the density distribution for information or resources, then optimal coverage solutions correspond to evenly distributed information or resources to each agent's "dominance region."

In [5], Cortés et al. design coverage control algorithms for a density-dependent metric defined, as a function of a given array of agents $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, by

$$\Phi(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n) = \int_{\mathcal{D}} \min_i \Big\{ f(d(\mathbf{x},\mathbf{x}_i))\, \rho(\mathbf{x}) \Big\}\, \mathrm{d}\mathbf{x},$$

where f is a nondecreasing function and ρ is the distribution density function. Because the metric depends on the Euclidean distance function, the cost function can be rewritten as

$$\Phi(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n) = \sum_{i=1}^{n} \int_{V_i} f(d(\mathbf{x},\mathbf{x}_i))\, \rho(\mathbf{x})\, \mathrm{d}\mathbf{x},$$

where the Voronoi cells V_i are defined also by the Euclidean distance function as

$$V_i = \left\{ \mathbf{x} \in \mathcal{D} \;\middle|\; d(\mathbf{x}, \mathbf{x}_i) \leq d(\mathbf{x}, \mathbf{x}_j) \ \forall j \neq i \right\}.$$

As shown in [5], this means that the cost function can be seen as the contribution of n dominance regions V_i, each of which is the Voronoi cell of an agent. Although this metric yields coverage solutions that are nonuniform, the information or resource will nonetheless *not* be equally distributed among corresponding dominance regions.

In this paper, we are interested in cost functions of the form (3.1). Indeed, a metric based on a nonuniform distance d_ρ is more closely related to information gathering and sensing array optimization.

One such problem is the detection of acoustic signals. In this case, $\sqrt{\rho}$ represents the nonuniform refractive index of the environment. The objective is to place the sensors in such a way that they can detect sources anywhere. In other words, one needs to minimize the weighted distance d_ρ between any point in the domain and the agents.

Another typical problem consists in increasing the ability of the array on an uneven terrain. This situation is typical for mine hunting arrays in a standby mode; the optimal configuration minimizes the time that it would take to send one of the agents to a newly detected mine. In this case, the square root of $\rho(\mathbf{x})$ represents the roughness of the terrain, the infinitesimal time it takes to cross an infinitesimal path located in $\mathbf{x}$. The goal is to position the agents in such a way that any point of the domain can be reached by one of the agents in minimum time. The optimal solution corresponds to the minimum of the cost function in (3.1), where $d_\rho(\mathbf{x}, \mathbf{y})$ is the minimum travel time between points $\mathbf{x}$ and $\mathbf{y}$.

A variant of the algorithm of [4, 5] could be used to optimize the coverage with a nonuniform metric defined in terms of a non-Euclidean distance function. Indeed, one can define Voronoi cells based on the non-Euclidean distance function. The boundaries of such Voronoi cells are, however, not polygonal, and their computation is complex and time-consuming. To compute the distance between two points $\mathbf{a}$ and $\mathbf{b}$, one needs to consider any path between $\mathbf{a}$ and $\mathbf{b}$ and to find the minimum of $\int_{\mathcal{C}_{\mathbf{a}}^{\mathbf{b}}} \sqrt{\rho}\, \mathrm{d}l$. Our approach, which involves computing cartograms, is, as will be illustrated below, a much simpler and faster operation.

In this paper, we assume that a particular cost function of the form (3.1) has been selected and that there exists a stable algorithm that brings a group of vehicles to the minimum of Φ for the Euclidean distance. We provide a methodology to modify this algorithm when the non-Euclidean distance (e.g., terrain roughness, acoustic refraction) is used.

4. Cartograms. Our approach to deriving coverage control strategies for nonuniform and time-varying metrics is to find a standard method to modify a control law defined for the Euclidean metric in such a way that it remains stable and converges to the minimum of the nonuniform metric.

The method that we develop is based on a nonuniform change of coordinates that transforms the domain $\mathcal{D}$ with the non-Euclidean distance into another compact set $\mathcal{D}'$ where the distance is Euclidean or near Euclidean. Such transformations are commonly referred to as “cartograms” in computer graphics.

To motivate the notion of cartogram, consider how poorly census and election results are represented using standard geographical projections; such data are better plotted on maps in which the sizes of geographic regions such as countries or provinces appear in proportion to their population (as opposed to the geographical area). Such maps, which are cartograms, transform the physical space $\mathcal{D}$ into a fictitious space $\mathcal{D}'$ where the area element A is proportional to a nonuniform density $\rho : \mathcal{D} \to \mathbb{R}_0^+$.

Definition 4.1 (cartogram). *Given a compact domain $\mathcal{D} \subset \mathbb{R}^2$ and a density function $\rho : \mathcal{D} \to \mathbb{R}_0^+$, a cartogram is a C^1 (continuous everywhere and with continuous derivatives almost everywhere) mapping $\phi : \mathcal{D} \to \mathcal{D}' : \mathbf{x} \to \phi(\mathbf{x})$ such that*

$$\det\left(\frac{\partial\phi}{\partial\mathbf{x}}\right) = \rho.$$

It is standard in the literature to define a cartogram as a change of coordinates as above but such that $\det\bigl(\frac{\partial\phi}{\partial\mathbf{x}}(\mathbf{x})\bigr) = k\rho(\mathbf{x}) > 0$; i.e., the transformation multiplies the area element by the density function ρ and an arbitrary constant $k > 0$. Without loss of generality, we assume that $k = 1$. If that is not the case, then one can multiply ϕ by k^{-1}.

As an example, Figure 2 shows the linguistic distribution in Belgium. The left panel of Figure 2 shows the five Flemish-speaking provinces and the five French-speaking provinces on an equal-area projection (Belgian conic conformal Lambert projection). The center panel gives the level sets of the population density and reveals that, while geographically smaller than its French counterpart, the Flemish region is much more densely populated and accounts for the majority of the country's population.

The right panel of Figure 2 is the cartogram of the country based on population density. In this projection, areas are proportional to the density of population. Such a cartogram is more adequate for plotting census and election results since the national outcome of the election or referendum is based on the principle of "one vote per citizen" and not "one vote per unit of area." Our method for computing cartograms, inspired by the approach of Gastner and Newman [9], is presented in this section.

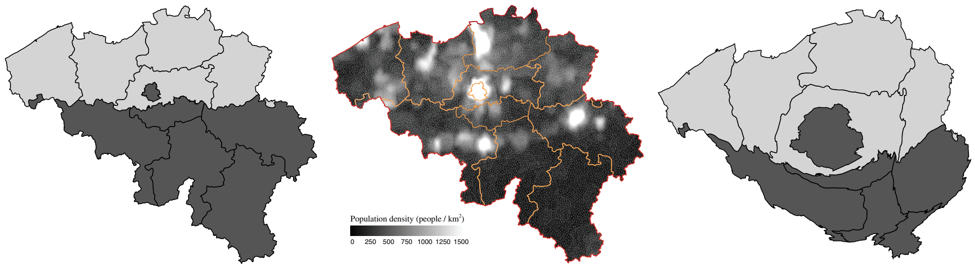


Fig. 2. *Linguistic distribution and population density in Belgium.* Left panel: *Equal-area (Belgian Lambert) projection. Light gray represents Flemish provinces. Dark gray stands for French-speaking provinces and the Brussels-Capital area.* Center panel: *Density of population binned on* 5km × 5km *cells (source: Columbia University's Center for International Earth Science Information Network).* Right panel: *Cartogram of the country based on population density. Areas in this projection are proportional to population density and correctly depict the linguistic distribution.*

Given a domain $\mathcal{D}$ and a density function ρ, there are infinitely many possible cartograms. As stated in [9], the objective is to minimize the distortion of the original figure. A perfect cartogram would not introduce any deformation and would satisfy

$$\frac{\partial\phi}{\partial\mathbf{x}} = \sqrt{\rho}\,\mathbb{I},$$

where $\mathbb{I}$ is the identity matrix. Clearly, such a cartogram does not exist for most density functions ρ. Nevertheless, we seek to reduce the distortion and to minimize $\left\|\frac{\partial\phi}{\partial\mathbf{x}} - \sqrt{\rho}\,\mathbb{I}\right\|$, where $\|\cdot\|$ is any norm on the space of 2×2 matrices. Accordingly, we make the following definitions.

Definition 4.2 (perfect cartogram). *For a given density function ρ, a perfect cartogram, if it exists, is a cartogram such that $\left\|\frac{\partial\phi}{\partial\mathbf{x}} - \sqrt{\rho}\,\mathbb{I}\right\| = 0$.*

Definition 4.3 (ideal cartogram). *For a given density function ρ, an ideal cartogram is given by*

$$\operatorname*{Argmin}_{\phi}\left\|\frac{\partial\phi}{\partial\mathbf{x}} - \sqrt{\rho}\,\mathbb{I}\right\|.$$

4.1. Cartograms with fixed boundaries. Recently, Gastner and Newman [9] showed how to construct a cartogram using a diffusion equation. Although perfect cartograms usually do not exist and there is no guarantee that a cartogram obtained using diffusion is an ideal cartogram, the work of Gastner and Newman has shown that, among all known methods to compute cartograms, the diffusion method introduces very little distortion and produces maps that are the closest to the perfect diagonal form $\sqrt{\rho}\,\mathbb{I}$.

To describe the method of [9], we first address the case in which, for a given ρ, the normal component of $\nabla\rho$ along the boundary $\partial\mathcal{D}$ vanishes. In this case, there exists a cartogram $\phi : \mathcal{D} \to \mathcal{D}'$, where $\mathcal{D}' = \mathcal{D}$. To show the existence of the cartogram and to determine a method to compute it, we imagine that the domain $\mathcal{D}$ is filled with a fluid whose initial density is given by ρ. As time evolves, the gradient of the density creates motion and the density of the fluid tends to homogenize. Let us consider the density $c(\mathbf{x}, t)$ at point $\mathbf{x}$ and time t. It satisfies the diffusion equation

$$\frac{\partial c}{\partial t} = \nu\Delta c\ ,$$

where the initial condition is $c(\mathbf{x}, 0) = \rho(\mathbf{x})$, the boundary condition is $\frac{\partial c}{\partial n} = 0$, and $\nu > 0$ is arbitrary. For $t \to +\infty$, the density c tends to a constant distribution c_∞ and, at any time t and at any position $\mathbf{x}$, we have $c(\mathbf{x}, t) > 0$. As a result, we can define a velocity field:

$$\mathbf{v}(\mathbf{x}, t) = -\nu\,\frac{\nabla c}{c}(\mathbf{x}, t).$$

Given the initial position $\mathbf{x}_0$ at which a particle is released at time $t = 0$, the velocity field above determines the position at any later time t. The flow (i.e., the trajectories) of the velocity field is a function of time and of the initial position. We denote by $\mathbf{x}(t; \mathbf{x}_0)$ the unique trajectory that satisfies

$$\begin{cases} \dot{\mathbf{x}} = \mathbf{v}\left(\mathbf{x}(t; \mathbf{x}_0), t\right), \\ \mathbf{x}(0; \mathbf{x}_0) = \mathbf{x}_0. \end{cases}$$

The domain $\mathcal{D}$ is compact; hence the trajectories $\mathbf{x}$ are at least C^1 on any *finite interval* of time $[0, t]$ (see, e.g., [12]). In this case, however, c is the solution of the diffusion equation and the magnitude of its gradient decays exponentially with time while c approaches its average, c_∞. As a result, the velocity field $\mathbf{v}$ also decays exponentially in time. This is a sufficient condition for the trajectories $\mathbf{x}(t, \mathbf{x})$ to be C^1 on the *infinite* interval $t \in [0, +\infty[$. We define

$$\phi(\mathbf{x}_0) = \lim_{t \to +\infty} \mathbf{x}(t, \mathbf{x}_0).$$

The limit exists, is unique, and is a C^1 function of its argument $\mathbf{x}_0$. To check that this transformation is indeed a cartogram, recall that Liouville's theorem determines how area elements A change along trajectories:

$$\frac{\mathrm{d}}{\mathrm{d}t} \ln A \bigg|_{\mathbf{x}(t;\mathbf{x}_0),t} = \operatorname{div}\left(\mathbf{v}(\mathbf{x}, t)\right). \tag{4.1}$$

Direct computation shows that

$$\operatorname{div}(\mathbf{v}) = -\frac{\nu}{c} \Delta c + \frac{1}{\nu} \mathbf{v}^2.$$

Note that

$$\frac{\mathrm{d}}{\mathrm{d}t} \ln c = \frac{1}{c} \frac{\partial c}{\partial t} + \frac{\mathbf{v} \cdot \nabla c}{c} = -\operatorname{div}(\mathbf{v}).$$

As a result, Liouville's equation (4.1) simplifies to

$$A(t) = A(0)\, e^{-\int_0^t \frac{\mathrm{d}}{\mathrm{d}t} \ln c \, \mathrm{d}t} = A(0)\, \frac{c(\mathbf{x}_0, 0)}{c(\mathbf{x}, t)} = A(0)\, \frac{\rho(\mathbf{x}_0)}{c(\mathbf{x}, t)}.$$

For $t \to +\infty$, the density c becomes constant in space and we have

$$\det\left(\frac{\partial \phi}{\partial \mathbf{x}_0}(\mathbf{x}_0)\right) = \lim_{t \to +\infty} \frac{A(t)}{A(0)} = \frac{\rho(\mathbf{x}_0)}{c_\infty}.$$

Hence, the transformation $\phi(\mathbf{x}_0)$ is a cartogram since it changes area elements proportionally to $\rho(\mathbf{x}_0)$. Starting from an equal-area map (i.e., $A(0) = 1$), we can multiply $\phi(\mathbf{x}_0)$ by the constant c_∞ and the equality becomes

$$\det\left(\frac{\partial \phi}{\partial \mathbf{x}_0}(\mathbf{x}_0)\right) = \rho(\mathbf{x}_0).$$

By changing the name of the variable from $\mathbf{x}_0$ to $\mathbf{x}$, we obtain the desired form

$$\det\left(\frac{\partial \phi}{\partial \mathbf{x}}(\mathbf{x})\right) = \rho(\mathbf{x}).$$

4.2. Cartograms with moving boundaries. The conclusions reached for cartograms with constant boundaries do not translate immediately to cases where $\frac{\partial \rho}{\partial n} \neq 0$ on the boundary of the domain. In this case, we cannot apply the method described in the previous section, and theorems about existence and smoothness of the diffusion problem, as well as about the advection of the velocity field, are not applicable either. Gastner and Newman [9] suggest extending the density to a larger domain where Neumann boundary conditions are enforced. Given a function $\rho : \mathcal{D} \to \mathbb{R}_0^+$, one can select a larger domain $\mathcal{D}_0 \supset \mathcal{D}$ and pick an arbitrary function $\hat{\rho} : \mathcal{D}_0 \to \mathbb{R}_0^+$ such that $\hat{\rho}$ is identical to ρ in $\mathcal{D}$. Typically, $\mathcal{D}_0$ has an area of 4 or 9 times the initial domain $\mathcal{D}$. The goal is to design $\hat{\rho}$ in such a way that $\frac{\partial \hat{\rho}}{\partial n} = 0$ at the edges of the larger domain $\mathcal{D}_0$

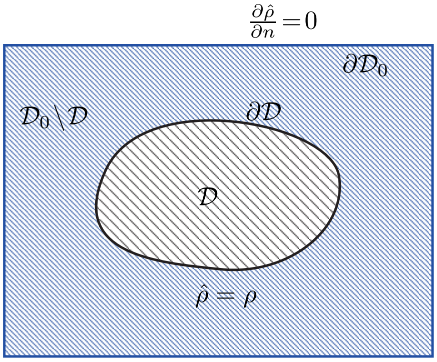

$\frac{\partial\hat\rho}{\partial n}=0$
$\partial\mathcal{D}_0$
$\mathcal{D}_0\setminus\mathcal{D}$
$\partial\mathcal{D}$
$\mathcal{D}$
$\hat\rho=\rho$

Fig. 3. *Proposed approach: when computing a cartogram for a domain $\mathcal{D}$ that has an arbitrary shape or for which the normal derivative of the density function ρ does not vanish at the boundary, a large rectangle $\mathcal{D}_0 \supset \mathcal{D}$ is selected. The density ρ is extended outside $\mathcal{D}$ by enforcing Neumann boundary conditions at the boundary of the large rectangle, requiring continuity of $\hat\rho$ at the edge with $\mathcal{D}$, and setting the Laplacian of $\hat\rho$ to a constant value outside $\mathcal{D}$. This defines a unique extension $\hat\rho$ which is continuous and has continuous derivatives almost everywhere.*

(see Figure 3). This permits the computation of the diffusion cartogram for the large domain $\mathcal{D}_0$ with fixed boundaries, followed by a restriction of the transformation to $\mathcal{D}$ to obtain the cartogram for the initial domain. This procedure is dependent on the choice of the embedding domain $\mathcal{D}_0$. Given $\mathcal{D}_0$, it also depends on how the extended density function $\hat\rho$ is constructed in $\mathcal{D}_0 \setminus \mathcal{D}$.

Gastner and Newman showed the importance of applying a "neutral buoyancy" condition, which keeps the total area under consideration constant. To construct $\hat\rho$, they first computed the average density in $\mathcal{D}$. In $\mathcal{D}_0 \setminus \mathcal{D}$, they filled $\hat\rho$ with a constant equal to the mean density in $\mathcal{D}$. They experimented with other choices of parameters, e.g., setting $\hat\rho = 0$ in $\mathcal{D}_0 \setminus \mathcal{D}$, but this resulted in inappropriate diffusion of density out of $\mathcal{D}$. The authors also experimented with different sizes for the domain $\mathcal{D}_0$ and observed only little visual difference, provided that $\mathcal{D}_0$ was sufficiently large.

From the point of view of our control design problem, the method above has an important flaw: $\hat\rho$, the initial condition for the diffusion problem, is not continuous at the boundary of $\mathcal{D}$. As a result, existence, uniqueness, and smoothness of the solution of the diffusion problem are not guaranteed. This is not necessarily a problem when producing only one cartogram. Our objective, however, is to produce continuous sequences of maps. Indeed, we will need the cartogram to vary smoothly when the density function is changed. For example, transferring Lyapunov functions from the cartogram space to the physical plane requires the existence of continuous derivatives.

As an alternative to Gastner and Newman's method, we propose the following variant. Given ρ in the domain of interest $\mathcal{D}$, we compute $\frac{\partial\rho}{\partial n}$ at the boundary of $\mathcal{D}$ and the total flux across $\partial\mathcal{D}$. We define the extended density $\hat\rho$ as follows:

- Inside $\mathcal{D}$, $\hat\rho(\mathbf{x}) = \rho(\mathbf{x})$.
- Outside $\mathcal{D}$, $\hat\rho$ is the solution of

$$\begin{cases} \Delta\hat\rho = \dfrac{-1}{\operatorname{Area}(\mathcal{D}_0\setminus\mathcal{D})}\displaystyle\int_{\mathcal{D}}\Delta\rho(\mathbf{x})\,\mathrm{d}\mathbf{x} = \dfrac{-1}{\operatorname{Area}(\mathcal{D}_0\setminus\mathcal{D})}\displaystyle\oint_{\partial\mathcal{D}}\frac{\partial\rho}{\partial n}\,\mathrm{d}l, \\ \left.\dfrac{\partial\hat\rho}{\partial n}\right|_{\partial\mathcal{D}_0} = 0, \\ \hat\rho|_{\partial\mathcal{D}} = \rho|_{\partial\mathcal{D}}. \end{cases}$$

The equations above define $\hat{\rho}$ inside $\mathcal{D}_0 \setminus \mathcal{D}$ as the solution of a linear problem with inhomogeneous Neumann boundary conditions. The Laplacian of $\hat{\rho}$ is constant in $\mathcal{D}_0 \setminus \mathcal{D}$, and its value is set so it compensates exactly the flux through the inside hole $\mathcal{D}$. Indeed, Green's equality requires

$$\int_{\mathcal{D}_0} \Delta\hat{\rho}\, \mathrm{d}\mathbf{x} = \oint_{\partial\mathcal{D}_0} \frac{\partial\hat{\rho}}{\partial n}\, \mathrm{d}l = 0.$$

Since this problem is compatible, standard results in functional analysis [1, 6] guarantee that the solution is unique and belongs to the Sobolev space H_1, which contains the functions on $\mathcal{D}_0$ that are continuous everywhere and for which the derivatives are continuous almost everywhere.[1] This guarantees also that the resulting extended density, $\hat{\rho}$, can be used as the initial condition of the diffusion problem and provides a C^1 solution c. The resulting transformation $\phi(\mathbf{x})$ is unique and varies smoothly (i.e., in a C^1 fashion) when the input density ρ is changed.

The only possible inconvenience is the fact that the solution $\hat{\rho}$ may become negative in the embedding rectangle. In this case the solution $\phi(\mathbf{x})$ might not be continuous due to the factor $\nabla c/c$ in the equation giving the velocity field. Nevertheless, the diffusion equation shows that if the initial density $c(\mathbf{x}, 0) = \hat{\rho}(\mathbf{x})$ is positive, then the density at later times remains positive. In other words, the density remain positive for all of the points that initially had a positive density. While the cartogram might not be well defined for points which were initially located in regions where $\hat{\rho}$ was negative, it is always well defined for points of $\mathcal{D}$ where the initial density is the input function $\rho > 0$.

4.3. Numerical methods. Gastner and Newman showed how the diffusion problem on a rectangle can be efficiently solved using the Fourier transform of $c(\mathbf{x}, t)$. This transforms the problem into an ordinary differential equation where the variables are the Fourier coefficients [9]. The only difference between our procedure and that of Gastner and Newman is how the density ρ is extended from the domain of interest $\mathcal{D}$ to the larger square $\mathcal{D}_0$. The problem giving $\hat{\rho}$ is linear; hence we can mesh $D_0 \setminus D$ and use a Galerkin approximation of $\hat{\rho}$ (see [1, 6]). Figure 4 illustrates the computation of $\hat{\rho}$ in $\mathcal{D}_0 \setminus \mathcal{D}$ for the domain in Figure 2. To compute the cartogram in the right panel of Figure 2, the extended density $\hat{\rho}$ in Figure 4 was used to first derive a map of the large rectangle. This is a necessary step, as we do not expect the normal derivative of the population density (input) to vanish at the border of a country.

Note that the partial differential equation that we suggest to solve for determining the extended density $\hat{\rho}$ is linear. The time needed to solve the linear problem is therefore negligible with respect to the time that it would take to compute the nonlinear boundaries of the Voronoi cells for the non-Euclidean metric. Another advantage of the procedure used here is that there already exist many optimized linear algebra packages that can be used to compute directly the solution of the discretized differential equation.

5. Nonuniform coverage control.

5.1. Method. Cartograms can be used to extend any algorithm that minimizes the uniform coverage metric, based on the Euclidean distance, to an algorithm that minimizes a nonuniform coverage metric dependent on an arbitrary "weighted" distance d_ρ. Indeed, starting from a non-Euclidean distance d_ρ, a perfect cartogram

[1] By "almost everywhere" we mean that the derivatives are continuous everywhere, except, possibly, on sets of measure zero.

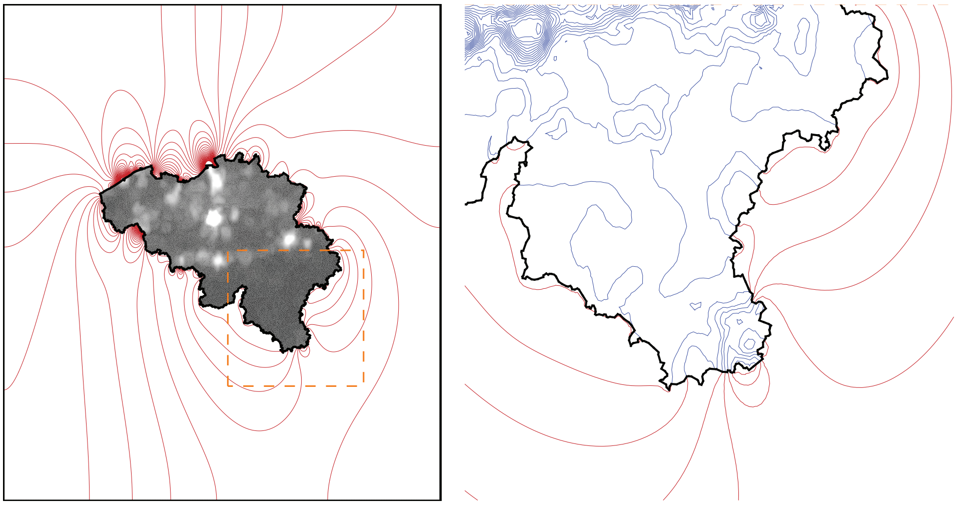


FIG. 4. *Continuous extension of the Belgian population density to a large rectangle with homogeneous Neumann boundary conditions. The extended density in the large rectangle is diffused to obtain the cartogram in the right panel of Figure* 2.

gives a transformation $\mathbf{y} = \phi(\mathbf{x})$ such that the distance function is Euclidean for $\mathbf{y}$. As a result, one can apply the uniform coverage algorithm to the $\mathbf{y}$ coordinates and prove convergence in the transformed space. In Theorem 5.1 (see section 5.3), we prove conditions under which convergence to the minimum of the uniform metric in the transformed space implies convergence to the minimum of the nonuniform metric in the original domain $\mathcal{D}$. The control law in the physical space for a system of agents with dynamics given by (2.1) can then be recovered from the chain rule as

$$\mathbf{u} = \dot{\mathbf{x}} = \left.\frac{\partial \phi^{-1}}{\partial \mathbf{y}}\right|_{\phi(\mathbf{x})} \dot{\mathbf{y}}.$$

5.2. Example. As an example, we let $\mathcal{D}$ be the unit square that was uniformly covered in Figure 1 using Cortés and Bullo's algorithm [4]. This time, however, we consider the multicenter coverage metric (2.2), where we replace the Euclidean distance d with a non-Euclidean distance d_ρ:

$$\Phi[d_\rho](\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = \max_{\mathbf{x}\in\mathcal{D}} \left\{ \min_{i=1\cdots n} d_\rho(\mathbf{x}, \mathbf{x}_i) \right\}. \tag{5.1}$$

For this example, we set the density function $\rho : \mathcal{D} \to \mathbb{R}_0^+$ to

$$\rho(x, y) = \frac{3}{40} + e^{-\frac{\left(x-\frac{3}{4}\right)^2 + \left(y-\frac{1}{4}\right)^2}{\left(\frac{1}{10}\right)^2}},$$

which represents our nonuniform interest in the features contained inside the unit square, or the nonuniform roughness of the terrain. The density ρ is plotted in the left panel of Figure 5. In this example, the lower right quadrant of the unit square has a much higher density and must be covered more densely than the rest of the square. In the analogy with a group of animals, the peak at $\left(\frac{3}{4}, \frac{1}{4}\right)$ represents a region with larger food supply and the population concentrates more in the lower right quadrant.

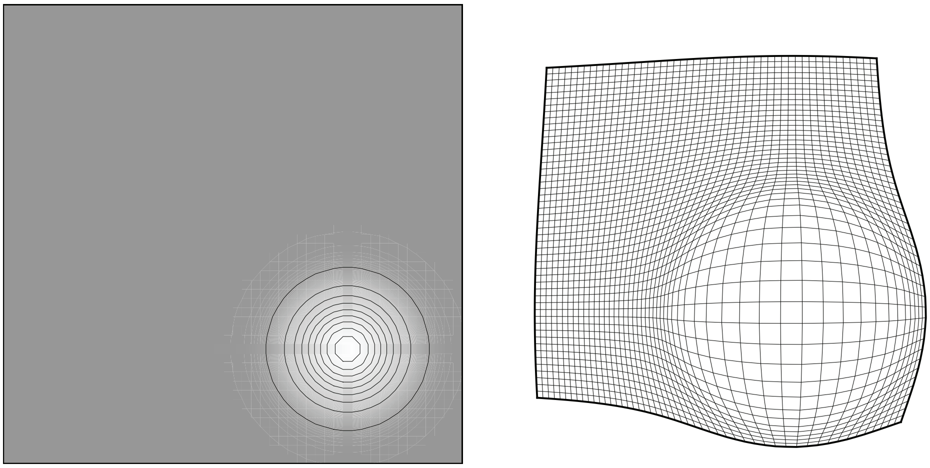


FIG. 5. *Cartogram of the unit square in preparation for nonuniform sampling.* Left panel: *Physical domain $\mathcal{D}$ with level sets of the density function $\rho(x,y) = 0.075 + \exp\left(-\frac{(x-0.75)^2+(y-0.25)^2}{0.1^2}\right)$.* Right panel: *Cartogram $\mathcal{D}'$ and image of a Cartesian mesh.*

In the analogy with a mine hunting array, the peak is a region where the robots move more slowly. To be able to respond anywhere in a minimum time, the vehicles must be closer to each other in the lower right quadrant.

To derive coverage control laws for $n = 4$ vehicles with dynamics given by (2.1), we first perform a cartogram of the area (see the right panel of Figure 5). The area near the peak of the Gaussian source is stretched by the transformation and represents about 30% of the mapped domain $\mathcal{D}'$, while it does not account for more than 10% of the physical domain $\mathcal{D}$.

To continue the example we apply the uniform coverage control law of [4] in the transformed plane $\mathcal{D}'$ which guarantees convergence to the optimal configuration in $\mathcal{D}'$. Figure 6 shows several snapshots of the motion of the particles in both $\mathcal{D}$ and $\mathcal{D}'$. Notice that the optimal configuration segments $\mathcal{D}'$ into four Voronoi cells of equal area, but, in the physical space, this corresponds to four (nonpolygonal) regions of unequal area; i.e., coverage is increased in the lower right quadrant.

Remark 1. Since the control is applied in a fictitious space, the norm of $\dot{\mathbf{y}}$ does not represent the actual speed of the vehicle. If unit speed control is desired (as in the simulation shown in Figure 6), then the velocity vector $\dot{\mathbf{y}}$ must be mapped back into the physical plane and normalized there.

Remark 2. The velocity in the cartogram is oriented along the segment between the vehicle and the circumcenter of its Voronoi cell. The preimage of this segment by the transformation is no longer a straight path. As a result, when the velocity vector is mapped back into physical coordinates by ϕ^{-1}, it is not necessarily oriented from the vehicle to the preimage of the circumcenter. Nevertheless, when the vehicle is infinitesimally close to the circumcenter, the two directions are aligned. At the equilibrium, the vehicles in the physical space are located exactly on the preimage of the circumcenter in the cartogram.

5.3. Proof of convergence. In this section, we prove that, given a feedback control law converging to a unique minimum of a cost function for the Euclidean metric, a perfect (or near perfect) cartogram provides a feedback control law that

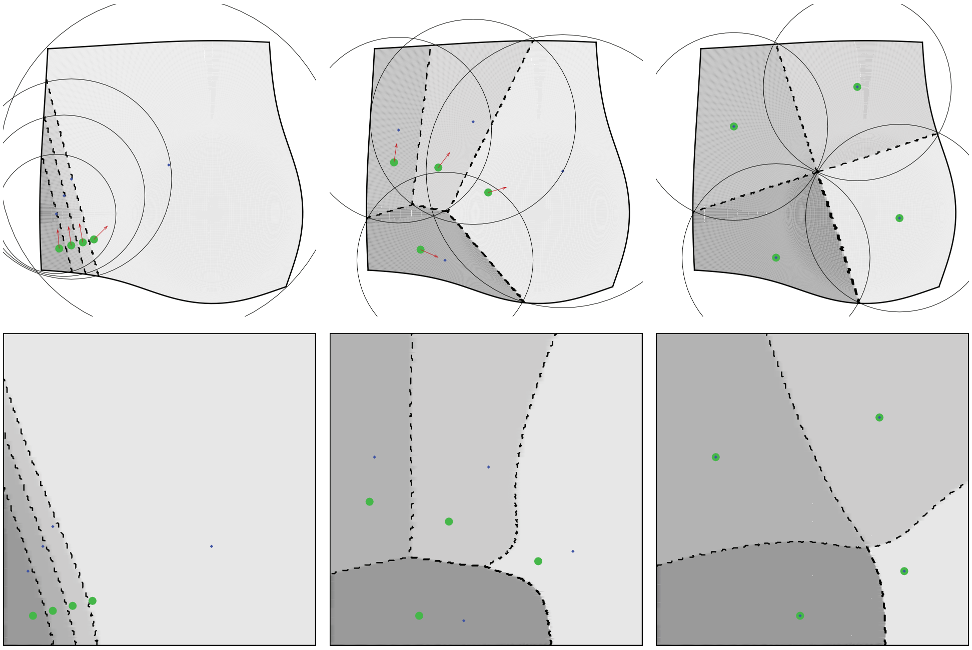


FIG. 6. *Convergence to static nonuniform coverage.* Thick dots: *Position of the four vehicles.* Shaded polygons: *Voronoi cell for each vehicle (computed in the cartogram space).* Large circles: *Circumcircle for each Voronoi cell of the cartogram.* Diamonds: *Centers of each circumcircle (i.e., circumcenters).* Arrows: *Instantaneous velocity of the vehicles (oriented along the segment joining the vehicle to the circumcenter of its Voronoi cell). The first row depicts the computation in the cartogram space. The second row gives the resulting positions of the vehicles in the physical space.*

converges to the unique minimum of the cost function for a non-Euclidean distance d_ρ defined by a density function $\rho > 0$.

Assume that a feedback control law has been designed and converges to the unique minimum of a cost function based on the Euclidean distance. We consider a nonuniform distance d_ρ and investigate how the control law for the Euclidean distance behaves in a near perfect cartogram of ρ. We show that, for C^1, strictly positive ρ, the non-Euclidean cost function has a unique minimum. Furthermore, the cartogram inverse-mapped feedback control converges toward this minimum.

THEOREM 5.1 (nonuniform coverage by cartograms). *Consider a C^1 cost function $\left(\Phi[d_\rho]\right)(\mathbf{x}_i, \ldots, \mathbf{x}_n; \mathcal{D})$ that depends only on the distance $d_\rho(\mathbf{a}, \mathbf{b}) = \min_{\mathcal{C}_{\mathbf{a}}^{\mathbf{b}}} \int_{\mathcal{C}_{\mathbf{a}}^{\mathbf{b}}} \sqrt{\rho}\,\mathrm{d}l$ between n agent positions and points in the domain $\mathcal{D}$. We assume that Φ has a unique, nondegenerate minimum for the Euclidean distance $d_1(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\|$. We also assume that there exists a feedback control law $\dot{\mathbf{x}}_i = \mathbf{v}_i(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ that brings the vehicles to the minimum* for the Euclidean distance d_1.

Given a density function $\rho : \mathcal{D} \to \mathbb{R}_0^+$, consider a cartogram $\phi : \mathcal{D} \to \phi(\mathcal{D})$. We consider applying the control law for the Euclidean distance in the cartogram space; hence

$$\dot{\mathbf{y}}_i = \mathbf{v}_i(\mathbf{y}_1, \ldots, \mathbf{y}_n),$$

where $\mathbf{y}_i = \phi(\mathbf{x}_i)$. *The corresponding dynamics in the physical space* $\mathcal{D}$

$$\dot{\mathbf{x}}_i = \mathbf{u}_i = \left.\frac{\partial \phi^{-1}}{\partial \mathbf{y}}\right|_{\phi(\mathbf{x})} \dot{\mathbf{y}}_i$$

yield a convergent sequence. In the neighborhood of a perfect cartogram, we have the following:

1. *There is a unique minimum of* $(\Phi[d_\rho])(\mathbf{x}_i, \dots, \mathbf{x}_n, \mathcal{D})$ *on* $\mathcal{D}$.
2. *The agents converge to an equilibrium that tends continuously to the unique minimum as* $\epsilon = \left\| \frac{\partial \phi}{\partial \mathbf{x}} - \sqrt{\rho}\, \mathbb{I} \right\| \to 0$.

Proof. By definition, a cartogram is a C^1 mapping

$$\phi : \mathcal{D} \to \phi(\mathcal{D}) : \mathbf{x} \mapsto \mathbf{y} = \phi(\mathbf{x}),$$

such that

$$\det\left(\frac{\partial \phi}{\partial \mathbf{x}}\right) = \rho.$$

This does not guarantee that the distance d_ρ between two points in the physical plane is equivalent to the Euclidean distance between the image of the two points in the cartogram. Nevertheless, the objective in computing the cartogram is to avoid unnecessary deformations, and, in the limit of a perfect cartogram, the equality of the distances is satisfied.

Consider a cartogram ϕ. The perfect cartogram is such that $\frac{\partial \phi}{\partial \mathbf{x}} = \sqrt{\rho}\, \mathbb{I}$, where $\mathbb{I}$ is the identity matrix. In section 6, we will show that the diffusion method described in this paper and in that of Gastner and Newman [9] provides a cartogram close to the ideal case for the examples that we studied. Developing the transformation about the perfect case, we get

$$\frac{\partial \phi}{\partial \mathbf{x}}(\mathbf{x}) = \sqrt{\rho}\, \mathbb{I} + \epsilon M(\mathbf{x}),$$

where $\|M\| = 1$ and $\epsilon = \left\| \frac{\partial \phi}{\partial \mathbf{x}} - \sqrt{\rho}\, \mathbb{I} \right\|$ is kept as small as possible to avoid distortions. For a perfect cartogram, we have $\epsilon \to 0$.

To investigate the relationship between distances in the physical space and in the cartogram, let us consider two points $\mathbf{x}_1$, $\mathbf{x}_2 \in \mathcal{D}$ and their images $\mathbf{y}_i = \phi(\mathbf{x}_i)$. We have

$$d_1(\mathbf{y}_1, \mathbf{y}_2) = \min_{\mathcal{C}_{\mathbf{y}_1}^{\mathbf{y}_2}} \int_{\mathcal{C}_{\mathbf{y}_1}^{\mathbf{y}_2}} \mathrm{d}l,$$

where $\mathcal{C}_{\mathbf{y}_1}^{\mathbf{y}_2}$ is an arbitrary path between points $\mathbf{y}_1$ and $\mathbf{y}_2$. Indeed, the Euclidean distance between two points $\mathbf{y}_1$ and $\mathbf{y}_2$ is the minimum length of the paths between the two points. Let us now apply the change of coordinates $\mathbf{y} = \phi(\mathbf{x})$. The infinitesimal arclength $\mathrm{d}l$ in the cartogram space becomes, in the physical space,

$$\sqrt{\mathbf{1}_l^\top \frac{\partial \phi}{\partial \mathbf{x}}^\top \frac{\partial \phi}{\partial \mathbf{x}} \mathbf{1}_l}\ \mathrm{d}l',$$

where $\mathbf{1}_l$ is the unit vector tangent to the path followed. As a result, we have

$$d_1(\mathbf{y}_1, \mathbf{y}_2) = \min_{\mathcal{C}_{\mathbf{y}_1}^{\mathbf{y}_2}} \int_{\mathcal{C}_{\mathbf{y}_1}^{\mathbf{y}_2}} \mathrm{d}l = \min_{\mathcal{C}_{\mathbf{x}_1}^{\mathbf{x}_2}} \int_{\mathcal{C}_{\mathbf{x}_1}^{\mathbf{x}_2}} \sqrt{\mathbf{1}_l^\top \frac{\partial \phi}{\partial \mathbf{x}}^\top \frac{\partial \phi}{\partial \mathbf{x}} \mathbf{1}_l} \ \mathrm{d}l' \tag{5.2}$$

$$= \min_{\mathcal{C}_{\mathbf{x}_1}^{\mathbf{x}_2}} \int_{\mathcal{C}_{\mathbf{x}_1}^{\mathbf{x}_2}} \sqrt{\rho + \epsilon \sqrt{\rho}\, \mathbf{1}_l^\top (M^\top + M) \mathbf{1}_l + \mathcal{O}(\epsilon^2)} \ \mathrm{d}l' \tag{5.3}$$

$$= \min_{\mathcal{C}_{\mathbf{x}_1}^{\mathbf{x}_2}} \int_{\mathcal{C}_{\mathbf{x}_1}^{\mathbf{x}_2}} \left[\sqrt{\rho} + \frac{\epsilon}{2\sqrt{\rho}} \, \mathbf{1}_l^\top (M^\top + M) \mathbf{1}_l + \mathcal{O}(\epsilon^2) \right] \ \mathrm{d}l' \tag{5.4}$$

$$= d_\rho(\mathbf{x}_1, \mathbf{x}_2) + \epsilon\, g(\mathbf{x}_1, \mathbf{x}_2) + \mathcal{O}(\epsilon^2), \tag{5.5}$$

where $\sqrt{\rho}\,\mathbf{1}_l^\top (M^\top + M)\mathbf{1}_l$ is continuous almost everywhere since ϕ is C^1 almost everywhere by definition. Notice that this also implies that $g(\mathbf{x}_1, \mathbf{x}_2) = \frac{\epsilon}{2\sqrt{\rho}} \int_{\mathcal{C}_{\mathbf{x}_1}^{\mathbf{x}_2}} \mathbf{1}_l^\top (M^\top + M)\mathbf{1}_l \ \mathrm{d}l$ is C^1 almost everywhere in its two arguments $\mathbf{x}_1$ and $\mathbf{x}_2$. As a result, the equation above states that the distance d_ρ between two points $\mathbf{x}_1$ and $\mathbf{x}_2$ in the physical space is equal to the Euclidean distance between the image of these two points in the cartogram $\phi(\mathbf{x}_1)$ and $\phi(\mathbf{x}_2)$, plus a correction $\epsilon\, g$ that vanishes continuously as $\epsilon \to 0$.

Since the cost function depends only on the distance between pairs of agents, between pairs of points in the domain $\mathcal{D}$, and between an agent and points of $\mathcal{D}$, we also have

$$\begin{aligned} \left(\Phi[d_\rho]\right)(\mathbf{x}_1, \ldots, \mathbf{x}_n; \mathcal{D}) = {} & \left(\Phi[d_1]\right)(\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_n), \phi(\mathcal{D})) \\ & + \epsilon\, h(\mathbf{x}_1, \ldots, \mathbf{x}_n) + \mathcal{O}(\epsilon^2). \end{aligned} \tag{5.6}$$

Since g is a C^1 function of its arguments, h is also C^1 almost everywhere. In other words, when $\epsilon \to 0$, *the cost function for a configuration in physical space with a nonuniform metric tends in a C^1 fashion to the cost function for the mapped configuration in the cartogram but with the Euclidean distance.*

Let us consider the (unique) minimum of the cost function with the Euclidean distance:

$$\nabla_{\mathbf{y}} \left(\Phi[d_1]\right)(\mathbf{y}_1^*, \ldots, \mathbf{y}_n^*; \phi(\mathcal{D})) = \mathbf{0}.$$

Recall that, by hypothesis, this minimum is unique and nondegenerate, i.e.,

$$\det \left(\left. \frac{\partial^2 \Phi}{\partial \mathbf{y}^2} \right|_{\mathbf{y}^*} \right) \neq 0.$$

Furthermore, the hypotheses guarantee convergence of the control in the cartogram for the Euclidean distance; hence the agents converge to the configuration $\mathbf{y}^*$. Since ϕ and ϕ^{-1} are continuous, the vehicles in the physical plane converge to $\mathbf{x}_i^* = \phi^{-1}(\mathbf{y}_i^*)$.

For a perfect cartogram, using the chain rule, we find

$$\nabla_{\mathbf{x}} \left(\Phi[d_\rho]\right)(\mathbf{x}_1, \ldots, \mathbf{x}_n; \mathcal{D}) = \underbrace{\left(\frac{\partial \phi}{\partial \mathbf{x}}\right)^{-1}}_{\det > 0} \nabla_{\mathbf{y}} \left(\Phi[d_1]\right)(\mathbf{y}_1, \ldots, \mathbf{y}_n; \phi(\mathcal{D})) .$$

Hence there is also one and only one minimum in the physical plane, and it is given by $\mathbf{x}_i^* = \phi^{-1}(\mathbf{y}_i^*)$. For a perfect cartogram, the vehicles converge exactly to the minimum of $\Phi[d_\rho]$.

For $\epsilon \neq 0$, the agents still reach the configuration $\mathbf{y}_i^*$ in the cartogram space. In the physical space, this still corresponds to the configuration $\mathbf{x}_i^* = \phi^{-1}(\mathbf{y}_i^*)$. In this case, however, $\mathbf{x}_i^*$ is not the minimum of $\Phi[d_\rho]$, and our goal is to show that the minimum of $\Phi[d_\rho]$ is still unique and that $\mathbf{x}_i^*$ is ϵ-close to this minimum. Starting from (5.6), we find

$$\begin{aligned}\nabla_{\mathbf{x}} \left(\Phi[d_\rho]\right)(\mathbf{x}_1,\ldots,\mathbf{x}_n;\mathcal{D}) &= \left(\frac{\partial\phi}{\partial\mathbf{x}}\right)^{-1} \nabla_{\mathbf{y}} \left(\Phi[d_1]\right)(\mathbf{y}_1,\ldots,\mathbf{y}_n;\phi(\mathcal{D})) \\ &\quad + \epsilon\, \nabla_{\mathbf{x}} h(\mathbf{x}_1,\ldots,\mathbf{x}_n) + \mathcal{O}(\epsilon^2),\end{aligned}$$

where $\nabla_{\mathbf{x}} h$ is continuous almost everywhere. Hence, the (possibly multiply defined) minima of $\Phi[d_\rho]$ satisfy

$$\nabla_{\mathbf{y}} \left(\Phi[d_1]\right)(\mathbf{y}_1,\ldots,\mathbf{y}_n;\phi(\mathcal{D})) = -\epsilon\, \frac{\partial\phi}{\partial\mathbf{x}}\, \nabla_{\mathbf{x}} h(\mathbf{x}_1,\ldots,\mathbf{x}_n) + \mathcal{O}(\epsilon^2).$$

For $\epsilon = 0$, the unique solution is $\mathbf{y}^*$. For ϵ sufficiently small, our goal is to show that the minimum, seen as a function $\mathbf{y}^*(\epsilon)$, is single-valued and continuous in ϵ. This is precisely the scope of the implicit function theorem (see, e.g., [19]); this classical result states that, unless the derivative of the relation versus $\mathbf{y}$ is singular at $(\mathbf{y},\epsilon) = (\mathbf{y}^*, 0)$, there exists such a C^1 curve of unique minima in a neighborhood of $\epsilon = 0$. Notice that the derivative of the implicit relationship at $\mathbf{y}^*$ and $\epsilon = 0$ is given by $\frac{\partial^2\Phi}{\partial\mathbf{y}^2}\big|_{\mathbf{y}^*}$, which is nonsingular by assumption. As a result, the implicit function theorem guarantees that, for sufficiently small ϵ, there exists a unique minimum $\mathbf{y}(\epsilon)$ and, as $\epsilon \to 0$, this minimum converges in a C^1 fashion to $\mathbf{y}^*$.

Now that we know that the minimum is unique and varies smoothly with ϵ, we can find its location. Setting $\mathbf{y}(\epsilon) = \mathbf{y}^* + \epsilon\,\boldsymbol{\delta}$, we find

$$\mathbf{y}_i^\epsilon = \mathbf{y}_i^* - \epsilon \left.\left(\frac{\partial^2\Phi}{\partial\mathbf{y}^2}\right)^{-1}\right|_{\mathbf{y}_i^*} \left.\frac{\partial\phi}{\partial\mathbf{x}}\right|_{\phi^{-1}(\mathbf{y}_i^*)} \nabla_{\mathbf{x}} h(\phi^{-1}(\mathbf{y}_i^*)) + \mathcal{O}(\epsilon^2).$$

This equation gives the coordinates (in the cartogram) of the minimum of $\Phi[d_\rho]$ (in physical space). The corresponding configuration in physical space is given by

$$\begin{aligned}\mathbf{x}_i^\epsilon &= \phi^{-1}(\mathbf{y}_i(\epsilon)) \\ &= \mathbf{x}_i^* - \epsilon \left.\left(\frac{\partial\phi}{\partial\mathbf{x}}\right)^{-1}\right|_{\mathbf{x}_i^*} \left.\left(\frac{\partial^2\Phi}{\partial\mathbf{y}^2}\right)^{-1}\right|_{\phi^{-1}(\mathbf{y}_i^*)} \left.\frac{\partial\phi}{\partial\mathbf{x}}\right|_{\mathbf{x}_i^*} \nabla_{\mathbf{x}} h(\mathbf{x}_i^*) + \mathcal{O}(\epsilon^2). \qquad (5.7)\end{aligned}$$

Notice that all of the matrices in the equation above are nonsingular. The minimum of $\Phi[d_\rho]$ is therefore also unique in the physical space and converges continuously to $\mathbf{x}^*$ as $\epsilon \to 0$.

Recall that $\mathbf{x}^*$ is the configuration to which the agents converge. $\mathbf{x}_i^\epsilon$ is the actual minimum of $\Phi[d_\rho]$. The equation above shows that, for small but nonzero ϵ, the optimum of the cost function with the nonuniform distance d_ρ is ϵ-close to the optimum of the cost function in the cartogram with the Euclidean metric. Furthermore, as $\epsilon \to 0$, the configuration of the agents converges continuously to the optimal configuration of the initial, non-Euclidean problem. $\square$

6. Cartogram error. We have shown in the previous section that, given two points $\mathbf{x}_1$ and $\mathbf{x}_2$ in space and their images by the cartogram, $\phi(\mathbf{x}_1)$ and $\phi(\mathbf{x}_1)$, we have

$$d_1\left(\phi(\mathbf{x}_1), \phi(\mathbf{x}_2)\right) \leq d_\rho(\mathbf{x}_1, \mathbf{x}_2) + \frac{\epsilon}{2\sqrt{\rho}} \left\|M^\top + M\right\| + \mathcal{O}\left(\epsilon^2\right).$$

In the context of controlling vehicles in the cartogram space, convergence is guaranteed, provided that we can create a cartogram such that $d_1\left(\phi(\mathbf{x}_1), \phi(\mathbf{x}_2)\right)$ is sufficiently close to $d_\rho(\mathbf{x}_1, \mathbf{x}_2)$. Recall that $\epsilon = \left\|\sqrt{\rho}\,\mathbb{I} - \frac{\partial \phi}{\partial \mathbf{x}}\right\|$ and $\|M\| = 1$; hence,

$$\left|\frac{\epsilon}{2\sqrt{\rho}}\right| \left\|M^\top + M\right\| \leq \left\|\mathbb{I} - \frac{1}{\sqrt{\rho}} \frac{\partial \phi}{\partial \mathbf{x}}\right\| \doteq \eta$$

is the relevant unitless distortion factor for this problem.

Figure 7 shows the distribution of the deformation factor η for the test cartogram shown in Figure 5 as well as for the cartogram based on the Belgian population density. In both cases, the maximum deformation factor η is below 0.4. Figure 7 shows also that the deformation is below 0.1 in a very high fraction of the total area. The distortion factor η grows above 0.1 only in small bands separating regions of very different densities. For example, there is a narrow annulus of high distortion around the density peak in the lower right corner of the left panel of Figure 7.

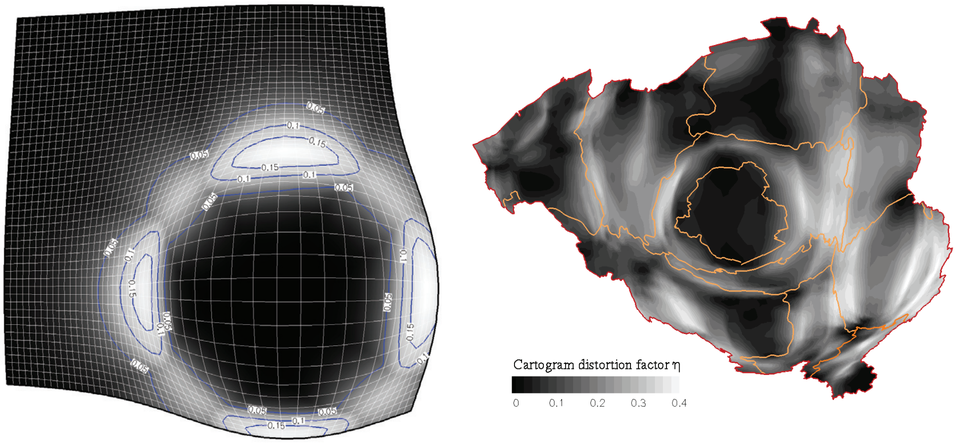


FIG. 7. *Level sets of the distortion factor* $\eta = \left\|I - \frac{1}{\rho}\frac{\partial \phi}{\partial \mathbf{x}}\right\|$ *for the cartograms constructed in Figure* 5 *(left panel) and in Figure* 2 *(right panel).*

It is worth noting that (5.7) indicates that we can always multiply the cost function Φ by an arbitrary large number to reduce the influence of the distortion factor. In the case of the example of section 5.2, the error can be estimated as follows: we have $\rho < 1.075$ and Figure 7 shows that, at the equilibrium positions, we have $\eta < 0.1$. As a result we have $\epsilon = \left\|\sqrt{\rho} I - \frac{\partial \phi}{\partial \mathbf{x}}\right\| = \sqrt{\rho}\,\eta < 0.11$. The error is given by $\frac{\epsilon}{2\sqrt{\rho}} = \frac{\eta}{2} < 0.05$, which is to be compared to the unit side of the square domain.

7. Space-time optimal coverage. In most problems, the metric does not remain constant in time. Indeed, when the density ρ is a physical quantity, such as the refractive index, it changes according to the fluctuations in the environment (e.g.,

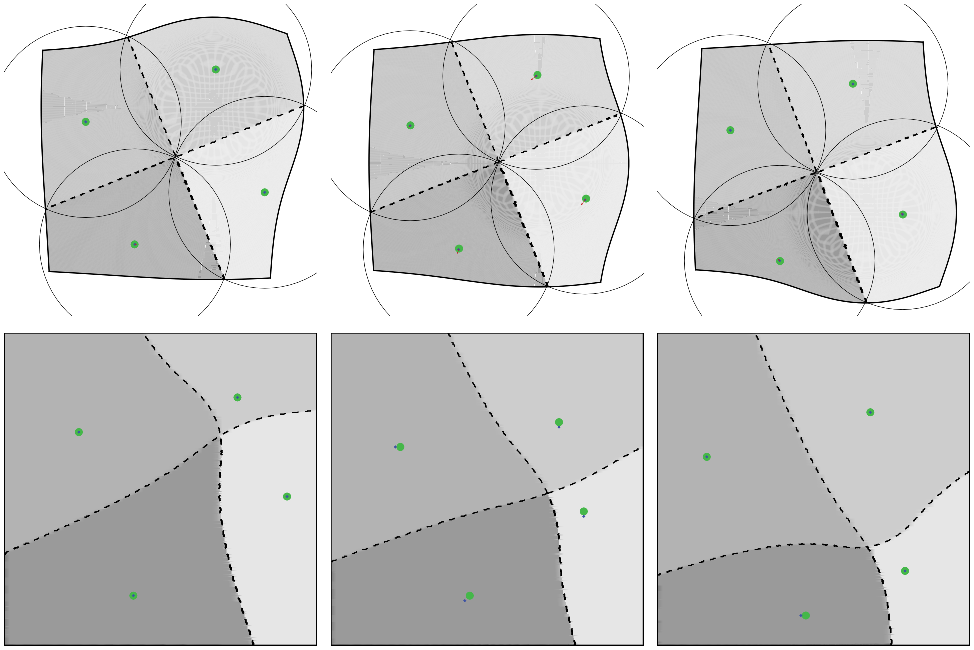


FIG. 8. *Convergence to nonuniform coverage with time-varying density.* Thick dots: *Position of the four vehicles.* Shaded polygons: *Voronoi cell for each vehicle (computed in the cartogram space).* Large circles: *Circumcircles for each Voronoi cell of the cartogram.* Diamonds: *Centers of each circumcircle (i.e., circumcenters).* Arrows: *Instantaneous velocity of the vehicles (oriented along the segment joining the vehicle to the circumcenter of its Voronoi cell). The first row depicts the computation in the cartogram space. The second row gives the resulting positions of the vehicles in the physical space. From left to right, the snapshots are taken when the peak of density is in the upper right corner, at $y = \frac{1}{2}$, and in the lower right corner.*

sources, sinks, diffusion, advection). When the nonuniform density represents information or food, its variations over time are even more subtle: the agents cover the domain and concentrate in regions of high density, but, at the same time, they deplete food or gather information and, in doing so, erode the very density peaks that attracted them. One such example is the objective analysis (OA) information map in [15].

The method developed in this paper is well suited for such time-varying metrics. Indeed, the numerical method presented in section 4 is aimed at producing cartograms that depend smoothly on the density function ρ. In other words, for a density function $\rho(\mathbf{x}, t)$ that is C^1 in time, we find a family of cartograms $\phi_t : \mathcal{D} \to \mathcal{D}'_t$ where both the transformation ϕ_t and the transformed space $\mathcal{D}'_t$ change with time in a C^1 fashion.

As an example, we modify the density function ρ from section 5.1 as follows:

$$\rho(x, y) = \frac{3}{40} + e^{-\frac{\left(x-\frac{3}{4}\right)^2 + \left(y - \frac{1}{2} + \frac{1}{4}\cos\left(\frac{t}{2\pi}\right)\right)^2}{\frac{1}{10}^2}}.$$

In other words, the peak of the distribution ρ now moves periodically along the vertical axis $x = \frac{3}{4}$. Figure 8 shows the result of this simulation for four agents. For the snapshots on the left, the peak of density is at its maximum position in the upper right quadrant. For the snapshots on the right, the peak is at the minimum in the lower right quadrant. The middle panels correspond to an intermediate position.

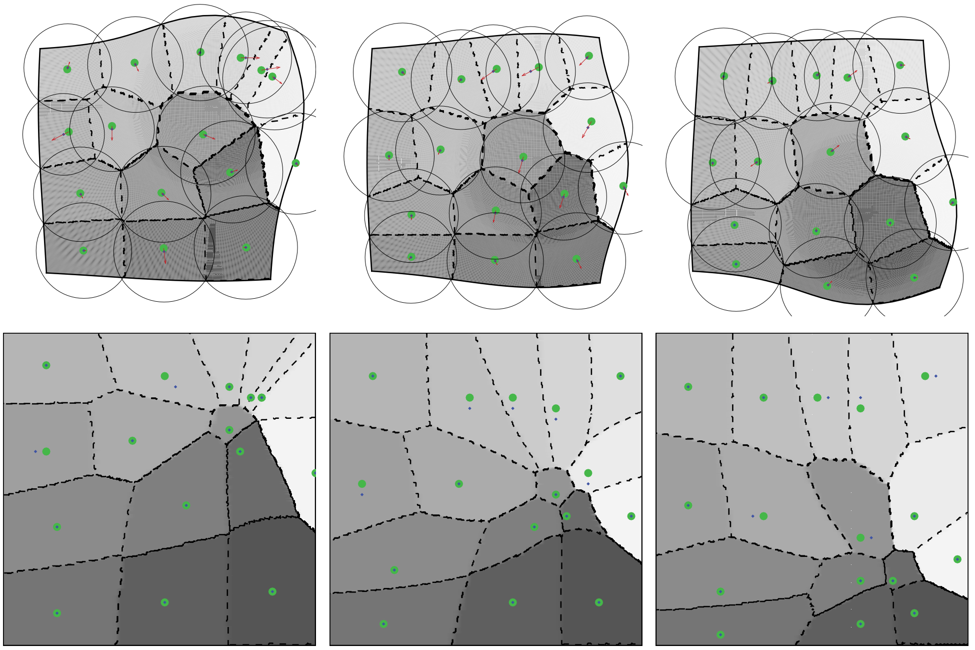


FIG. 9. *Array of* 10 *agents covering the square with nonuniform, time-varying density. On the left, the peak of density is in the upper right quadrant; on the right, the peak is in the lower right quadrant. The middle panel is an intermediate snapshot.*

Notice that, despite the fact that the optimal configuration changes a lot in physical space, the cartograms are similar to each other. This highlights one of the advantages of the method: the complexity of the nonuniform and time-varying density is absorbed in the cartogram transformation. In the cartogram plane, a simpler, uniform coverage algorithm is applied.

In Figure 9, the same simulation is performed for 10 vehicles, and it shows how the agents organize and move to follow the peak in the density ρ.

For autonomous, nonuniform metrics, we proved uniqueness of the optimal configuration and convergence to this position. The algorithm applies well to the case of time-varying metrics. If the density function changes slowly enough (in comparison to agent speed) and, at any time t, the distortion $\eta(t) = \left\| \mathbb{I} - \frac{1}{\sqrt{\rho(t)}} \frac{\partial \phi}{\partial \mathbf{x}}(t) \right\|$ is sufficiently small, then convergence can be inferred by our theorem. The requirement that $\rho(t)$ does not change too fast guarantees that the cartogram does not change too fast, and, hence, the boundary $\phi(\mathcal{D})$ does not change too fast with respect to vehicle speed. As a result, the motion of the cartogram boundary (slow dynamics) and the motion of the vehicles (fast dynamics) are almost decoupled, and we infer convergence from the fact that vehicles are converging to the equilibrium on timescales much shorter than the timescale at which $\phi(\mathcal{D})$ changes.

For a fast changing metric, there are two possible obstacles that can limit the use of the method that we proposed. First, a fast changing metric might make the cartogram boundary move too fast for the vehicle to converge to any configuration. Second, if the vehicles converge, then the configuration reached might differ significantly from the minimum of the metric.

8. Conclusions. In this paper, we investigated the use of cartograms to achieve time-varying, nonuniform coverage of a spatial domain by a group of agents. The method proposed relies on the existence of a stable algorithm that achieves uniform coverage (e.g., for the Euclidean metric). The control law is extended to nonuniform coverage (based on a non-Euclidean distance induced by a density function) by the use of density-equalizing maps.

The advantage of the method presented is its universality: it permits generalizing many existing uniform coverage algorithms to nonuniform metrics. It also provides a simple and fast control law. For example, computing Voronoi cells with a nonuniform metric is, in comparison, a very time-consuming operation.

The control law presented in this paper is not distributed. In particular, the diffusion equation uses information from all agents and is computed on a central computer. However, a distributed version of the proposed approach may be possible, provided that each agent computes its own local diffusion equation. In this case information passed from neighbors would be used to determine boundary conditions.

Another limitation of the method presented in this paper is the fact that it applies only to simple vehicle dynamics (first-order control). Although it is possible to find heuristic adaptations to nonholonomic constraints, uniqueness of the optimal configuration and convergence are not guaranteed by the results presented in this paper. The major difficulty is that nonholonomic constraints, unlike distances and cost functions, are not preserved by the cartogram mapping.

An inspiration for this work comes from a desire to understand how animal groups organize themselves to exploit a time-varying nonuniform food supply. Other applications are numerous, as teams of unmanned robots are often destined to tasks for which they need to mimic foraging animals. In [15], underwater vehicles patrol the ocean to collect scientific data. Information for these agents is analogous to food: the vehicles "consume" information by going to unsampled areas and lowering the uncertainty. After a region has been visited, uncertainty grows in time, analogous to the growing "food source."

OA provides a framework to study quantitatively these systems by estimating the residual error (or negative information) for an array of agents. It has been used successfully to optimize the design of static and moving arrays [3, 15]. The method presented in this paper is designed to match the OA model. Indeed, the residual error corresponds to the input density ρ. Adequate data sampling is a key ingredient in providing accurate ocean models. Recent developments in ocean modeling provide detailed error maps [16] that can be equally used as the input density of our method. Our objective is to optimize the motion of sampling agents in such a way that the residual error in the models assimilating these data is minimum. From this point of view, the cartograms represent an abstraction layer between a complex objective (minimize error in large scale ocean models) and the control algorithm itself.

Acknowledgments. The authors are grateful to Russ Davis (Scripps Institution of Oceanography), Pierre Lermusiaux (MIT), and David Fratantoni (Woods Hole Oceanographic Institution) for enlightening discussions about underwater glider control and ocean sampling. The Fourier transforms needed to solve the diffusion equation on rectangles were performed using FFTW, the "fastest Fourier transform in the West" [8]. Finite-element approximations of the linear partial differential equation problems defining the extended density function between the domain of interest and a larger embedding rectangle were obtained using libsparse, a C++ wrapper to the University of Florida's UMFPack (see documentation of the OMA package or [13, 14]).

REFERENCES

[1] J.-P. Aubin, *Applied Functional Analysis*, Pure Appl. Math. (N.Y.), Wiley-Interscience, New York, 2000.

[2] C. Bernstein, A. Kacelnik, and J. R. Krebs, *Individual decisions and the distribution of predators in a patchy environment*, J. Animal Ecology, 57 (1988), pp. 1007–1026.

[3] F. P. Bretherton, R. E. Davis, and C. B. Fandry, *A technique for objective analysis and design of oceanographic experiments applied to MODE*-73, Deep-Sea Res., 23 (1976), pp. 559–582.

[4] J. Cortés and F. Bullo, *Coordination and geometric optimization via distributed dynamical systems*, SIAM J. Control Optim., 44 (2005), pp. 1543–1574.

[5] J. Cortés, S. Martínez, T. Karatas, and F. Bullo, *Coverage control for mobile sensing networks*, IEEE Trans. Robotics and Automation, 20 (2004), pp. 243–255.

[6] B. Daya Reddy, *Introductory Functional Analysis*, Springer-Verlag, New York, 1997.

[7] S. D. Fretwell and H. L. Lucas, Jr., *On territorial behavior and other factors influencing habitat distribution in birds*, Acta Biotheoretica, 19 (1970), pp. 16–36.

[8] M. Frigo and S. G. Johnson, *The design and implementation of FFTW*3, Proc. IEEE, 93 (2005), pp. 216–231.

[9] M. T. Gastner and E. J. Newman, *Diffusion-based method for producing density-equalizing maps*, Proc. Natl. Acad. Sci. USA, 101 (2004), pp. 7499–7504.

[10] M. T. Gastner and E. J. Newman, *Optimal design of spatial distribution networks*, Phys. Rev. E (3), 74 (2006), 016117.

[11] M. T. Gastner, C. R. Shalizi, and E. J. Newman, *Maps and cartograms of the 2004 US presidential election results*, Adv. Complex Syst., 8 (2005), pp. 117–123.

[12] M. W. Hirsch and S. Smale, *Differential Equations, Dynamical Systems and Linear Algebra*, Academic Press, New York, 1974.

[13] D. M. Kaplan and F. Lekien, *Spatial interpolation and filtering of surface current data based on open-boundary modal analysis*, J. Geophys. Res. [Oceans], 112 (2007), C12007.

[14] F. Lekien, C. Coulliette, R. Bank, and J. Marsden, *Open-boundary modal analysis: Interpolation, extrapolation, and filtering*, J. Geophys. Res. [Oceans], 109 (2004), C12004.

[15] N. E. Leonard, D. Paley, F. Lekien, R. Sepulchre, D. M. Fratantoni, and R. Davis, *Collective motion, sensor networks and ocean sampling*, Proc. IEEE, 95 (2007), pp. 48–74.

[16] P. F. J. Lermusiaux, *Uncertainty estimation and prediction for interdisciplinary ocean dynamics*, J. Comput. Phys., 217 (2006), pp. 176–199.

[17] W. Li and C. G. Cassandras, *Distributed cooperative coverage control of sensor networks*, in Proceedings of the 44th IEEE Conference on Decision and Control, 2005, pp. 2542–2547.

[18] S. Poduri and G. S. Sukhatme, *Constrained coverage for mobile sensor networks*, in Proceedings of the IEEE International Conference on Robotics and Automation, 2004, pp. 165–172.

[19] W. Rudin, *Principles of Mathematical Analysis*, McGraw–Hill, New York, 1976.

[20] D. W. Stephens and J. R. Krebs, *Foraging Theory*, Princeton University Press, Princeton, NJ, 1986.

SIAM J. Control Optim.
Vol. 48, No. 1, pp. 373–396

© 2009 Society for Industrial and Applied Mathematics

PAYOFF-BASED DYNAMICS FOR MULTIPLAYER WEAKLY ACYCLIC GAMES*

JASON R. MARDEN[†], H. PEYTON YOUNG[‡], GÜRDAL ARSLAN[§], AND JEFF S. SHAMMA[¶]

Abstract. We consider repeated multiplayer games in which players repeatedly and simultaneously choose strategies from a finite set of available strategies according to some strategy adjustment process. We focus on the specific class of weakly acyclic games, which is particularly relevant for multiagent cooperative control problems. A strategy adjustment process determines how players select their strategies at any stage as a function of the information gathered over previous stages. Of particular interest are "payoff-based" processes in which, at any stage, players know only their own actions and (noise corrupted) payoffs from previous stages. In particular, players do not know the actions taken by other players and do not know the structural form of payoff functions. We introduce three different payoff-based processes for increasingly general scenarios and prove that, after a sufficiently large number of stages, player actions constitute a Nash equilibrium at any stage with arbitrarily high probability. We also show how to modify player utility functions through tolls and incentives in so-called congestion games, a special class of weakly acyclic games, to guarantee that a centralized objective can be realized as a Nash equilibrium. We illustrate the methods with a simulation of distributed routing over a network.

Key words. game theory, cooperative control, learning in games

AMS subject classifications. 91A10, 91A80, 68W15

DOI. 10.1137/070680199

1. Introduction.

The objective in distributed cooperative control for multiagent systems is to enable a collection of "self-interested" agents to achieve a desirable "collective" objective. There are two overriding challenges to achieving this objective. The first is complexity. Finding an optimal solution by a centralized algorithm may be prohibitively difficult when there are large numbers of interacting agents. This motivates the use of adaptive methods that enable agents to "self-organize" into suitable, if not optimal, collective solutions.

The second challenge is limited information. Agents may have limited knowledge about the status of other agents, except perhaps for a small subset of "neighboring" agents. An example is collective motion control for mobile sensor platforms (see, e.g., [7]). In these problems, mobile sensors seek to position themselves to achieve various collective objectives such as rendezvous or area coverage. Sensors can communicate with neighboring sensors, but otherwise they do not have global knowledge of the domain of operation or the status and locations of nonneighboring sensors.

A typical assumption is that agents are endowed with a reward or utility function

*Received by the editors January 15, 2007; accepted for publication (in revised form) October 14, 2008; published electronically February 11, 2009. This research was supported by NSF grant ECS-0501394, ARO grant W911NF–04–1–0316, and AFOSR grant FA9550-05-1-0239.
http://www.siam.org/journals/sicon/48-1/68019.html

[†] Information Science and Technology, California Institute of Technology, Pasadena, CA 91125 (marden@caltech.edu).

[‡] Department of Economics, University of Oxford and the Brookings Institute, Oxford OX1 3UQ, UK (pyoung@brookings.edu).

[§] Department of Electrical Engineering, University of Hawaii, Honolulu, HI 96822 (gurdal@hawaii.edu).

[¶] Corresponding author. School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 (shamma@gatech.edu).

that depends on their own strategies and the strategies of other agents. In motion coordination problems, for example, an agent's utility function typically depends on its position relative to other agents or environmental targets, and knowledge of this function guides local motion adjustments.

In other situations, agents may know nothing about the structure of their utility functions and how their own utility depends on the actions of other agents (whether local or far away). In this case, the only course of action is to observe rewards based on experience and "optimize" on a trial and error basis. The situation is further complicated because all agents are trying simultaneously to optimize their own strategies. Therefore, even in the absence of noise, an agent trying the same strategy twice may see different results because of the nonstationary nature of the strategies of other agents.

There are several examples of multiagent systems that illustrate this situation. In distributed routing for ad hoc data networks (see, e.g., [2]), routing nodes seek to route packets to neighboring nodes based on packet destinations without knowledge of the overall network structure. The objective is to minimize the delay of packets to their destinations. This delay must be realized through trial and error, since the functional dependence of delay on routing strategies is not known. A similar problem is automotive traffic routing, in which drivers seek to minimize the congestion experienced to reach a desired destination. Drivers can experience the congestion on selected routes as a function of the routes selected by other drivers, but drivers do not know the structure of the congestion function. Finally, in a multiagent approach to designing manufacturing systems (see, e.g., [9]), it may not be known in advance how performance measures (such as throughput) depend on manufacturing policy. Rather, performance can only be measured once a policy is implemented.

Our interest in this paper is to develop algorithms that enable coordination in multiagent systems for precisely this "payoff-based" scenario, in which agents only have access to (possibly noisy) measurements of the rewards received through repeated interactions with other agents. We adopt the framework of "learning in games." (See [5, 10, 25, 26] for an extensive overview. See also the recent special issue containing [22] or survey article [18] for perspectives from machine learning.) Unlike most of the learning rules in this literature, which assume that agents adjust their behavior based on the observed behavior of other agents, we shall assume that agents know only their own past actions and the payoffs that resulted. It is far from obvious that Nash equilibrium can be achieved under such a restriction, but in fact it has recently been shown that such "payoff-based" learning rules can be constructed that work in any game [4, 8].

In this paper we show that there are simpler and more intuitive adjustment rules that achieve this objective for a large class of multiplayer games known as "weakly acyclic" games. This class captures many problems of interest in cooperative control [13, 14]. It includes the very special case of "identical interest" games, where each agent receives the same reward. However, weakly acyclic games (and the related concept of potential games) capture other scenarios such as congestion games [19] and similar problems such as distributed routing in networks, weapon target assignment, consensus, and area coverage. See [15, 1] and references therein for a discussion of a learning in games approach to cooperative control problems, but under less stringent assumptions on informational constraints than considered in this paper.

For many multiagent problems, operation at a pure Nash equilibrium may reflect optimization of a collective objective.[1] We will derive payoff-based dynamics that

[1] Nonetheless, there are varied viewpoints on the role of Nash equilibrium as a solution concept for multiagent systems. See [22] and [12].

guarantee asymptotically that agent strategies will constitute a pure Nash equilibrium with arbitrarily high probability. It need not always be the case that at least one Nash equilibrium optimizes a collective objective. Motivated by this consideration, we also discuss the introduction of incentives or tolls in a player's payoff function to assure that there is at least one Nash equilibrium that optimizes a collective objective. Even in this case, however, there may still be suboptimal Nash equilibria.

The remainder of this paper is organized as follows. Section 2 provides background on finite strategic-form games and repeated games. This is followed by three types of payoff-based dynamics in section 3 for increasingly general problems. Subsection 3.1 presents "safe experimentation dynamics" which is restricted to identical interest games. Subsection 3.2 presents "simple experimentation dynamics" for the more general class of weakly acyclic games but with noise-free payoff measurements. Subsection 3.3 presents "sample experimentation dynamics" for weakly acyclic games with noisy payoff measurements. Section 4 discusses how to introduce tolls and incentives in payoffs so that a Nash equilibrium optimizes a collective objective. Section 5 presents an illustrative example of a traffic congestion game. Finally, section 6 contains some concluding remarks. An important analytical tool throughout is the method of resistance trees for perturbed Markov chains [24], which is reviewed in an appendix.

2. Background.

In this section, we will present a brief background of the game theoretic concepts used in the paper. We refer the readers to [6, 25, 26] for a more comprehensive review.

2.1. Finite strategic-form games.

Consider a finite strategic-form game with n-player set $\mathcal{P} := \{\mathcal{P}_1, \dots, \mathcal{P}_n\}$ where each player $\mathcal{P}_i \in \mathcal{P}$ has a finite action set $\mathcal{A}_i$ and a utility function $U_i : \mathcal{A} \to \mathbb{R}$ where $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$. We will sometimes use a single symbol, e.g., G, to represent the entire game, i.e., the player set, $\mathcal{P}$, action sets, $\mathcal{A}_i$, and utility functions U_i.

For an action profile $a = (a_1, a_2, \dots, a_n) \in \mathcal{A}$, let a_{-i} denote the profile of player actions *other than* player $\mathcal{P}_i$, i.e.,

$$a_{-i} = \{a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_n\}\,.$$

With this notation, we will sometimes write a profile a of actions as (a_i, a_{-i}). Similarly, we may write $U_i(a)$ as $U_i(a_i, a_{-i})$.

An action profile $a^* \in \mathcal{A}$ is called a *pure Nash equilibrium* if for all players $\mathcal{P}_i \in \mathcal{P}$,

$$U_i(a_i^*, a_{-i}^*) = \max_{a_i \in \mathcal{A}_i} U_i(a_i, a_{-i}^*). \tag{2.1}$$

Furthermore, if the above condition is satisfied with a unique maximizer for every player $\mathcal{P}_i \in \mathcal{P}$, then a^* is called a *strict (Nash) equilibrium*.

In this paper we will consider three classes of games: identical interest games, potential games, and weakly acyclic games. Each class of games has a connection to general cooperative control problems and multiagent systems for which there is some global utility or potential function $\phi : \mathcal{A} \to \mathbb{R}$ that a global planner seeks to maximize [13].

2.1.1. Identical interest games.

The most restrictive class of games that we will review in this paper is identical interest games. In such a game, the players' utility functions $\{U_i\}_{i=1}^n$ are chosen to be the same. That is, for some function $\phi : \mathcal{A} \to \mathbb{R}$,

$$U_i(a) = \phi(a)$$

for every $\mathcal{P}_i \in \mathcal{P}$ and for every $a \in \mathcal{A}$. It is easy to verify that all identical interest games have at least one pure Nash equilibrium, namely, any action profile a that maximizes $\phi(a)$.

2.1.2. Potential games. A significant generalization of an identical interest game is a potential game. In a potential game, the change in a player's utility that results from a unilateral change in strategy equals the change in the global utility. Specifically, there is a function $\phi : \mathcal{A} \to \mathbb{R}$ such that for every player $\mathcal{P}_i \in \mathcal{P}$, for every $a_{-i} \in \mathcal{A}_{-i}$, and for every $a_i', a_i'' \in \mathcal{A}_i$,

$$U_i(a_i', a_{-i}) - U_i(a_i'', a_{-i}) = \phi(a_i', a_{-i}) - \phi(a_i'', a_{-i}).$$

When this condition is satisfied, the game is called an exact potential game with the potential function ϕ.[2] It is easy to see that, in potential games, any action profile maximizing the potential function is a pure Nash equilibrium, and hence every potential game possesses at least one such equilibrium. An example of an exact potential game is illustrated in Figure 1.

	L	R
U	$0, 0$	$-1, 1$
D	$1, -1$	$0, 0$

Payoffs

	L	R
U	0	1
D	1	2

Potential

FIG. 1. *An example of a two player exact potential game.*

2.1.3. Weakly acyclic games. Consider any finite game G with a set $\mathcal{A}$ of action profiles. A *better reply path* is a sequence of action profiles $a^1, a^2, \ldots, a^L$ such that for each successive pair a^j, a^{j+1} there is exactly one player such that $a_i^j \neq a_i^{j+1}$ and for that player $U_i(a^{j+1}) > U_i(a^j)$. In other words, one player moves at a time, and each time a player moves he increases his own utility.

Suppose now that G is a potential game with potential function ϕ. Starting from an arbitrary action profile $a \in \mathcal{A}$, construct a better reply path $a = a^1, a^2, \ldots, a^L$ until it can no longer be extended. Note first that such a path cannot cycle back on itself, because ϕ is strictly increasing along the path. Since $\mathcal{A}$ is finite, the path cannot be extended indefinitely. Hence, the last element in a maximal better reply path from any joint action, a, must be a Nash equilibrium of G.

This idea may be generalized as follows. The game G is *weakly acyclic* if for any $a \in \mathcal{A}$, there exists a better reply path starting at a and ending at some pure Nash equilibrium of G [25, 26]. Potential games are special cases of weakly acyclic games. An example of a two player weakly acyclic game is illustrated in Figure 2. Notice that the illustrated game is not a potential game.

2.2. Repeated games. In a repeated game, at each time $t \in \{0, 1, 2, \ldots\}$, each player $\mathcal{P}_i \in \mathcal{P}$ simultaneously chooses an action $a_i(t) \in \mathcal{A}_i$ and receives the utility $U_i(a(t))$, where $a(t) := (a_1(t), \ldots, a_n(t))$. Each player $\mathcal{P}_i \in \mathcal{P}$ chooses action $a_i(t)$ at time t according to a probability distribution $p_i(t)$, which we will refer to as the

[2] There are weaker notions of potential games such as ordinal or weighted potential games. Rather than discuss each variation specifically, we will discuss a more general framework, weakly acyclic games, in the ensuing section. Any potential game, whether exact, ordinal, or weighted, is a weakly acyclic game.

	L	C	R
U	$0, 0$	$0.1, 0$	$1, 1$
M	$1, 0$	$0, 1$	$0, 0$
D	$0, 1$	$1, 0$	$0, 0$

FIG. 2. *An example of a two player weakly acyclic game.*

strategy of player $\mathcal{P}_i$ at time t. A player's strategy at time t can rely only on observations from times $\{0, 1, 2, \ldots, t-1\}$. Different learning algorithms are specified by both the assumptions on available information and the mechanism by which the strategies are updated as information is gathered. For example, if a player knows the functional form of his utility function and is capable of observing the actions of all other players at every time step, then the strategy adjustment mechanism of player $\mathcal{P}_i$ can be written in the general form

$$p_i(t) = F_i\big(a(0), \ldots, a(t-1); U_i\big).$$

An example of a learning algorithm, or strategy adjustment mechanism, of this form is the well-known fictitious play [16]. For a detailed review of learning in games, we direct the reader to [5, 25, 26, 11, 23, 20].

In this paper we deal with the issue of whether players can learn to play a pure Nash equilibrium through repeated interactions under the most restrictive observational conditions; players *only* have access to (i) the action they played and (ii) the utility (possibly noisy) they received. In this setting, the strategy adjustment mechanism of player $\mathcal{P}_i$ takes on the form

$$p_i(t) = F_i\big(\{a_i(0), U_i(a(0)) + \nu_i(0)\}, \ldots, \{a_i(t-1), U_i(a(t-1)) + \nu_i(t-1)\}\big), \tag{2.2}$$

where the $\nu_i(t)$ are zero mean independent and identically distributed (i.i.d.) random variables.

3. Payoff-based learning algorithms.

In this section, we will introduce three simple payoff-based learning algorithms. The first, called *safe experimentation*, guarantees convergence to a pure optimal Nash equilibrium in any identical interest game. Such an equilibrium is optimal because each player's utility is maximized. The second learning algorithm, called *simple experimentation*, guarantees convergence to a pure Nash equilibrium in any weakly acyclic game. The third learning algorithm, called *sample experimentation*, guarantees convergence to a pure Nash equilibrium in any weakly acyclic game even when utility measurements are corrupted with noise.

3.1. Safe experimentation dynamics for identical interest games.

3.1.1. Constant exploration rates. Before introducing the learning dynamics, we introduce the following function. Let

$$U_i^{\max}(t) := \max_{0 \le \tau \le t-1} U_i(a(\tau))$$

be the maximum utility that player $\mathcal{P}_i$ has received up to time $t-1$.

We will now introduce the safe experimentation dynamics for identical interest games.

1. **Initialization:** At time $t = 0$, each player randomly selects and plays any action, $a_i(0)$. This action will be initially set as the player's *baseline action* at time $t = 1$ and is denoted by $a_i^b(1) = a_i(0)$.

2. **Action selection:** At each subsequent time step, each player selects his baseline action with probability $(1-\epsilon)$ or experiments with a new random action with probability ϵ, i.e.,
   - $a_i(t) = a_i^b(t)$ with probability $(1-\epsilon)$;
   - $a_i(t)$ is chosen randomly (uniformly) over $\mathcal{A}_i$ with probability ϵ.

   The variable ϵ will be referred to as the player's *exploration rate*.
3. **Baseline strategy update:** Each player compares the actual utility received, $U_i(a(t))$, with the maximum received utility $U_i^{\max}(t)$ and updates the baseline action as follows:

$$a_i^b(t+1) = \begin{cases} a_i(t), & U_i(a(t)) > U_i^{\max}(t), \\ a_i^b(t), & U_i(a(t)) \le U_i^{\max}(t). \end{cases}$$

   Each player updates the maximum received utility regardless of whether or not step 2 involved exploration.
4. Return to step 2 and repeat.

The reason that this learning algorithm is called "safe" experimentation is that the utility evaluated at the baseline action, $U(a^b(t))$, is nondecreasing with respect to time.

THEOREM 3.1. *Let G be a finite n-player identical interest game in which all players use the safe experimentation dynamics. Given any probability $p < 1$, if the exploration rate $\epsilon > 0$ is sufficiently small, then for all sufficiently large times t, $a(t)$ is an optimal Nash equilibrium of G with at least probability p.*

Proof. Since G is an identical interest game, let the utility of each player be expressed as $U : \mathcal{A} \to \mathbb{R}$, and let $\mathcal{A}^*$ be the set of "optimal" Nash equilibria of G, i.e.,

$$\mathcal{A}^* = \left\{ a^* \in \mathcal{A} : U(a^*) = \max_{a \in \mathcal{A}} U(a) \right\}.$$

For any joint action, $a(t)$, the ensuing joint action will constitute an optimal Nash equilibrium with at least probability

$$\left(\frac{\epsilon}{|\mathcal{A}_1|}\right)\left(\frac{\epsilon}{|\mathcal{A}_2|}\right)\cdots\left(\frac{\epsilon}{|\mathcal{A}_n|}\right),$$

where $|\mathcal{A}_i|$ denotes the cardinality of the action set of player $\mathcal{P}_i$. Therefore, an optimal Nash equilibrium will eventually be played with probability 1 for any $\epsilon > 0$.

Suppose an optimal Nash equilibrium is first played at time t^*, i.e., $a(t^*) \in \mathcal{A}^*$ and $a(t^*-1) \notin \mathcal{A}^*$. Then the baseline joint action must remain constant from that time onwards, i.e., $a^b(t) = a(t^*)$ for all $t > t^*$. An optimal Nash equilibrium will then be played at any time $t > t^*$ with at least probability $(1-\epsilon)^n$. Since $\epsilon > 0$ can be chosen arbitrarily small, and in particular such that $(1-\epsilon)^n > p$, this completes the proof. ◻

3.1.2. Diminishing exploration rates. In the safe experimentation dynamics, the exploration rate ϵ was defined as a constant. Alternatively, one could let the exploration rate vary to induce desirable behavior. One example would be to let the exploration rate decay, such as $\epsilon_t = (1/t)^{1/n}$. This would induce exploration at early stages and reduce exploration at later stages of the game. The theorem and proof hold under the following conditions for the exploration rate:

$$\lim_{t\to\infty} \epsilon_t = 0,$$

$$\lim_{t\to\infty} \prod_{\tau=1}^{t} \left[1 - \left(\frac{\epsilon_\tau}{|\mathcal{A}_1|}\right)\left(\frac{\epsilon_\tau}{|\mathcal{A}_2|}\right) \cdots \left(\frac{\epsilon_\tau}{|\mathcal{A}_n|}\right) \right] = 0.$$

3.2. Simple experimentation dynamics for weakly acyclic games. We will now introduce the simple experimentation dynamics for weakly acyclic games. These dynamics will allow us to relax the assumption of identical interest games.

1. **Initialization:** At time $t = 0$, each player randomly selects and plays any action, $a_i(0)$. This action will be initially set as the player's *baseline action* at time 1, i.e., $a_i^b(1) = a_i(0)$. Likewise, the player's *baseline utility* at time 1 is initialized as $u_i^b(1) = U_i(a(0))$.
2. **Action selection:** At each subsequent time step, each player selects a baseline action with probability $(1 - \epsilon)$ or experiments with a new random action with probability ϵ, i.e.,
   - $a_i(t) = a_i^b(t)$ with probability $(1 - \epsilon)$;
   - $a_i(t)$ is chosen randomly (uniformly) over $\mathcal{A}_i$ with probability ϵ.

   The variable ϵ will be referred to as the player's *exploration rate*. Whenever $a_i(t) \neq a_i^b(t)$, we will say that player $\mathcal{P}_i$ *experimented*.
3. **Baseline action and baseline utility update:** Each player compares the utility received, $U_i(a(t))$, with his baseline utility, $u_i^b(t)$, and updates his baseline action and utility as follows:
   - If player $\mathcal{P}_i$ *experimented* (i.e., $a_i(t) \neq a_i^b(t)$) and if $U_i(a(t)) > u_i^b(t)$, then

     $a_i^b(t+1) = a_i(t)$,
     $u_i^b(t+1) = U_i(a(t))$.
   - If player $\mathcal{P}_i$ *experimented* and if $U_i(a(t)) \leq u_i^b(t)$, then

     $a_i^b(t+1) = a_i^b(t)$,
     $u_i^b(t+1) = u_i^b(t)$.
   - If player $\mathcal{P}_i$ *did not experiment* (i.e., $a_i(t) = a_i^b(t)$), then

     $a_i^b(t+1) = a_i^b(t)$,
     $u_i^b(t+1) = U_i(a(t))$.
4. Return to step 2 and repeat.

As before, these dynamics require only utility measurements and hence almost no information regarding the structure of the game.

THEOREM 3.2. *Let G be a finite n-player weakly acyclic game in which all players use the simple experimentation dynamics. Given any probability $p < 1$, if the exploration rate $\epsilon > 0$ is sufficiently small, then for all sufficiently large times t, $a(t)$ is a Nash equilibrium of G with at least probability p.*

The remainder of this subsection is devoted to the proof of Theorem 3.2. The proof relies on the theory of resistance trees for perturbed Markov chains (see the appendix for a brief review).

Define the *state* of the dynamics to be the pair $[a, u]$, where a is the baseline joint action and u is the baseline utility vector. We will omit the superscript b to avoid cumbersome notation.

Partition the state space into the following three sets. First, let X be the set of states $[a, u]$ such that $u_i \neq U_i(a)$ for at least one player $\mathcal{P}_i$. Let E be the set of states $[a, u]$ such that $u_i = U_i(a)$ for all players $\mathcal{P}_i$ and a is a Nash equilibrium. Let D be the set of states $[a, u]$ such that $u_i = U_i(a)$ for all players $\mathcal{P}_i$ and a is a disequilibrium (not a Nash equilibrium). These are all the states.

Claim 3.1.

(a) *Any state $[a, u] \in X$ transitions to a state in $E \cup D$ in one period with probability $O(1)$.*

(b) *Any state $[a, u] \in E \cup D$ transitions to a different state $[a', u']$ with probability at most $O(\varepsilon)$.*

Proof. For any $[a, u'] \in X$, there exists at least one player $\mathcal{P}_i$ such that $u'_i \neq U_i(a)$. If all players repeat their part of the joint action profile a, which occurs with probability $(1-\epsilon)^n$, then $[a, u']$ transitions to $[a, u]$, where $u_i = U_i(a)$ for all players $\mathcal{P}_i$. Thus the process moves to $[a, u] \in E \cup D$ with prob $O(1)$. This proves statement (a). As for statement (b), any state in $E \cup D$ transitions back to itself whenever no player experiments, which occurs with probability at least $O(1)$. □

Claim 3.2. *For any state $[a, u] \in D$, there is a finite sequence of transitions to a state $[a^*, u^*] \in E$, where the transitions have the form*[3]

$$[a, u] \underset{O(\epsilon)}{\rightarrow} [a^1, u^1] \underset{O(\epsilon)}{\rightarrow} \cdots \underset{O(\epsilon)}{\rightarrow} [a^*, u^*],$$

where $u_i^k = U_i(a^k)$ for all i and for all $k > 0$, and each transition occurs with probability $O(\epsilon)$.

Proof. Such a sequence is guaranteed by weak acyclicity. Since a is not an equilibrium, there is a better reply path from a to some equilibrium a^*, say $a, a^1, a^2, \ldots, a^*$.

At $[a, u]$ the appropriate player $\mathcal{P}_i$ experiments with probability ϵ and chooses the appropriate better reply with probability $1/|\mathcal{A}_i|$, and no one else experiments. Thus the process moves to $[a^1, u^1]$, where $u_i^1 = U_i(a^1)$ for all players $\mathcal{P}_i$ with probability $O(\epsilon)$ (more precisely, $O(\epsilon(1-\epsilon)^{n-1})$). Notice that for the deviator $\mathcal{P}_i$, $U_i(a^1) > U_i(a)$, and therefore $u_i^1 = U_i(a^1)$. For the nondeviator, say, player $\mathcal{P}_j$, $u_j^1 = U_j(a^1)$ since $a_j^1 = a_j$. Thus $[a^1, u^1] \in D \cup E$. In the next period, the appropriate player deviates, and so forth. □

Claim 3.3. *For any equilibrium $[a^*, u^*] \in E$, any path from $[a^*, u^*]$ to another state $[a, u] \in E \cup D$, $a \neq a^*$, that does not loop back to $[a^*, u^*]$ must be one of the following two forms:*

(1) $[a^*, u^*] \underset{O(\epsilon)}{\rightarrow} [a^*, u'] \underset{O(\epsilon^k)}{\rightarrow} [a', u''] \rightarrow \cdots \rightarrow [a, u]$, *where $k \geq 1$;*

(2) $[a^*, u^*] \underset{O(\epsilon^k)}{\rightarrow} [a', u''] \rightarrow \cdots \rightarrow [a, u]$, *where $k \geq 2$.*

Proof. The path must begin by either one player experimenting or more that one player experimenting. Case (2) results if more than one player experiments. Case (1) results if exactly one agent, say, agent $\mathcal{P}_i$, experiments with an action $a'_i \neq a^*_i$ and all other players continue to play their part of a^*. This happens with probability $(\epsilon/|\mathcal{A}_i|)(1-\epsilon)^{n-1}$. In this situation, player $\mathcal{P}_i$ cannot be better off, meaning that $U_i(a'_i, a^*_{-i}) \leq U_i(a^*)$, since by assumption a^* is an equilibrium. Hence the baseline action next period remains a^* for all players, though their baseline utilities may change. Denote the next state by $[a^*, u']$. If in the subsequent period all players continue to play their part of the action a^*, which occurs with probability $(1-\epsilon)^n$, then the state reverts back to $[a^*, u^*]$ and we have a loop. Hence, the only way the path can continue without a loop is for one or more players to experiment in the next stage, which has probability $O(\epsilon^k)$, $k \geq 1$. This is exactly what case (1) alleges. □

Proof of Theorem 3.2. This is a finite aperiodic Markov process on the state space $\mathcal{A} \times \bar{U}_1 \times \cdots \times \bar{U}_n$, where $\bar{U}_i$ denotes the (finite) range of $U_i(\cdot)$. Furthermore, from

[3] We will use the notation $z \rightarrow z'$ to denote the transition from state z to state z'. We use $z \underset{O(\epsilon)}{\rightarrow} z'$ to emphasize that this transition occurs with probability of order ϵ.

every state there exists a positive probability path to a Nash equilibrium. Hence, every recurrent class has at least one Nash equilibrium. We will now show that within any recurrent class, the trees (see the appendix) rooted at the Nash equilibrium will have the lowest resistance. Therefore, according to Theorem A.1, the a priori probability that the state will be a Nash equilibrium can be made arbitrarily close to 1.

In order to apply Theorem A.1, we will construct minimum resistance trees with vertices consisting of every possible state (within a recurrence class). Each edge will have resistance $0, 1, 2, \dots$ associated with the transition probabilities $O(1), O(\epsilon)$, $O(\epsilon^2), \dots$, respectively.

Our analysis will deviate slightly from the presentation in the appendix. In the discussion in the appendix, the vertices of minimum resistance trees are recurrence classes of an associated unperturbed Markov chain. In this case, the unperturbed Markov chain corresponds to simple experimentation dynamics with $\epsilon = 0$, and so the recurrence classes are all states in $E \cup D$. Nonetheless, we will construct resistance trees with the vertices being all possible states, i.e., $E \cup D \cup X$. The resulting conclusions remain the same (see Lemma 1 in [24]). Since the states in X are transient with probability $O(1)$, the resistance to leave a node corresponding to a state in X is 0. Therefore, the presence of such states does not affect the conclusions determining which states are stochastically stable.

Suppose a minimum resistance tree T is rooted at a vertex v that is not in E. If $v \in X$, it is easy to construct a new tree that has lower resistance. Namely, by Claim 3.1(a), there is a zero-resistance one-hop path P from v to some state $[a, u] \in E \cup D$. Add the edge of P to T and subtract the edge in T that exits from the vertex $[a, u]$. This results in a $[a, u]$-tree T'. It has lower resistance than T because the added edge has zero resistance, while the subtracted edge has resistance greater than or equal to 1 because of Claim 3.1(b). This argument is illustrated in Figure 3, where the edge of strictly positive resistance ($R \geq 1$) is removed and replaced with the edge of zero resistance ($R = 0$).

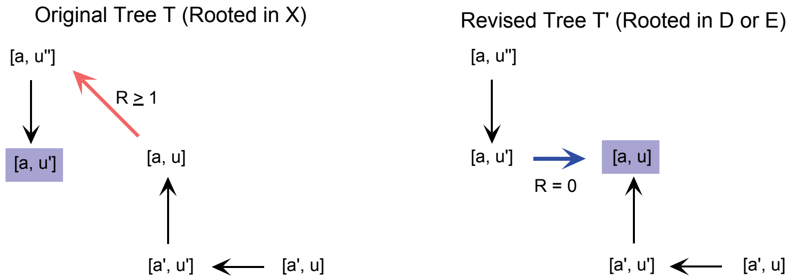


FIG. 3. *Construction of alternative to tree rooted in X.*

Suppose next that $v = [a, u] \in D$ but not in E. Construct a path P as in Claim 3.2 from $[a, u]$ to some state $[a^*, u^*] \in E$. As above, construct a new tree T' rooted at $[a^*, u^*]$ by adding the edges of P to T and taking out the redundant edges (the edges in T that exit from the vertices in P). The nature of the path P guarantees that the edges taken out have total resistance at least as high as the resistances of the edges put in. This is because the entire path P lies in $E \cup D$, each transition on the path has resistance 1, and, from Claim 3.2(b), the resistance to leave any state in $E \cup D$ is at least 1.

To construct a new tree that has strictly lower resistance, we will inspect the effect of removing the exiting edge from $[a^*, u^*]$ in T. Note that this edge must fit either case (1) or (2) of Claim 3.3.

In case (2), the resistance of the exiting edge is at least 2, which is larger than any edge in P. Hence the new tree has strictly lower resistance than T, which is a contradiction. This argument is illustrated in Figure 4. A new path is created from the original root $[a, u] \in D$ to the equilibrium $[a^*, u^*] \in E$ ($R = 1$ edges). Redundant ($R \geq 1$, $R \geq 2$) edges emanating from the new path are removed. In case (2), the redundant edge emanating from $[a^*, u^*]$ has a resistance of at least 2.

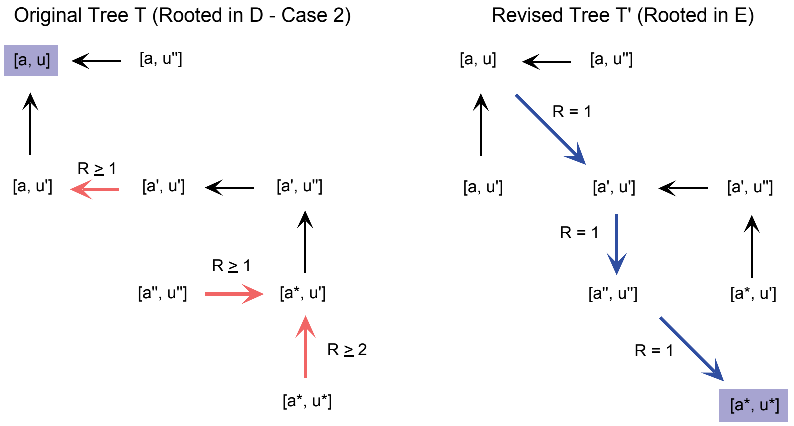


FIG. 4. *Construction of alternative to tree rooted in D for case* (2).

In case (1), the exiting edge has the form $[a^*, u^*] \to [a^*, u']$ which has resistance 1 where $u^* \neq u'$. The next edge in T, say, $[a^*, u'] \to [a', u'']$, also has at least resistance 1. Remove the edge $[a^*, u'] \to [a', u'']$ from T, and put in the edge $[a^*, u'] \to [a^*, u^*]$. The latter has resistance 0 since $[a^*, u'] \in X$. This results in a tree T'' that is rooted at $[a^*, u^*]$ and has strictly lower resistance than does T, which is a contradiction. This argument is illustrated in Figure 5. As in Figure 4, a new ($R = 1$, $R = 0$) path is constructed and redundant ($R \geq 1$, $R = 1$) edges are removed. The difference is that the edge $[a^*, u'] \to [a', u'']$ is removed and replaced with $[a^*, u'] \to [a^*, u^*]$.

To recap, a minimum resistance tree cannot be rooted at any state in X or D, but rather only at a state in in E. Therefore, when ϵ is sufficiently small, the long-run probability on E can be made arbitrarily close to 1, and in particular, larger than any specified probability p. $\square$

3.3. Sample experimentation dynamics for weakly acyclic games with noisy utility measurements.

3.3.1. Noise-free utility measurements. In this section we will focus on developing payoff-based dynamics for which the limiting behavior exhibits that of a pure Nash equilibrium with arbitrarily high probability in any finite weakly acyclic game *even in the presence of utility noise*. We will show that a variant of the so-called regret testing algorithm [4] accomplishes this objective for weakly acyclic games with noisy utility measurements.

We now introduce sample experimentation dynamics.

1. **Initialization:** At time $t = 0$, each player randomly selects and plays any action, $a_i(0) \in \mathcal{A}_i$. This action will be initially set as each player's *baseline action*, $a_i^b(1) = a_i(0)$.

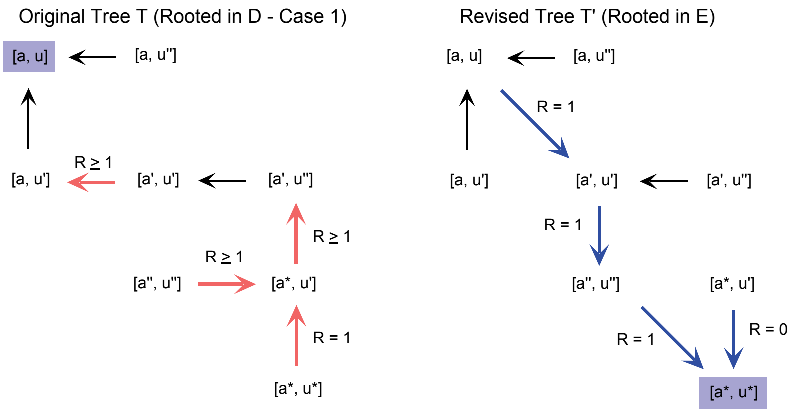


FIG. 5. *Construction of alternative to tree rooted in D for case* (1).

2. **Exploration phase:** After the baseline action is set, each player engages in an *exploration phase* over the next m periods. The exploration phases need not be synchronized or of the same length for each player, but we will assume that they are for the proof. For convenience, we will double index the time of the actions played as

$$\check{a}(t_1, t_2) = a(m\,t_1 + t_2),$$

where t_1 indexes the number of the exploration phase and t_2 indexes the actions played in that exploration phase. We will refer to t_1 as the *exploration phase time* and to t_2 as the *exploration action time*. By construction, the exploration phase time and exploration action time satisfy $t_1 \geq 1$ and $m \geq t_2 \geq 1$, respectively. The baseline action will be updated only at the end of the exploration phase and will therefore be indexed only by the exploration phase time.
During the exploration phase, each player selects a baseline action with probability $(1-\epsilon)$ or experiments with a new random action with probability ϵ. That is, for any exploration phase time $t_1 \geq 1$ and for any exploration action time satisfying $m \geq t_2 \geq 1$,
   - $\check{a}_i(t_1, t_2) = a_i^b(t_1)$ with probability $(1-\epsilon)$,
   - $\check{a}_i(t_1, t_2)$ is chosen randomly (uniformly) over $(\mathcal{A}_i \setminus a_i^b(t_1))$ with probability ϵ.

   Again, the variable ϵ will be referred to as the player's *exploration rate*.
3. **Action assessment:** After the exploration phase, each player evaluates the average utility received when playing each of his actions during the exploration phase. Let $n_i^{a_i}(t_1)$ be the number of times that player $\mathcal{P}_i$ played action a_i during the exploration phase at time t_1. The average utility for action a_i during the exploration phase at time t_1 is

$$\hat{V}_i^{a_i}(t_1) = \begin{cases} \frac{1}{n_i^{a_i}(t_1)} \sum_{t_2=1}^{m} I\{a_i = \check{a}_i(t_1,t_2)\} U_i(\check{a}(t_1,t_2)), & n_i^{a_i}(t_1) > 0, \\ U_{\min}, & n_i^{a_i}(t_1) = 0, \end{cases}$$

where $I\{\cdot\}$ is the usual indicator function and $U_{\min}$ satisfies

$$U_{\min} < \min_i \min_{a \in \mathcal{A}} U_i(a).$$

In other words, $U_{\min}$ is less than the smallest payoff any agent can receive.

4. **Evaluation of better response set:** Each player compares the average utility received when playing a baseline action, $\hat{V}_i^{a_i^b(t)}(t_1)$, with the average utility received for each of the other actions, $\hat{V}_i^{a_i}(t_1)$, and finds all played actions which performed δ better than the baseline action. The term δ will be referred to as the players' *tolerance level*. Define $\mathcal{A}_i^*(t_1)$ to be the set of actions that outperformed the baseline action as follows:

$$\mathcal{A}_i^*(t_1) = \left\{ a_i \in \mathcal{A}_i : \hat{V}_i^{a_i}(t_1) \geq \hat{V}_i^{a_i^b(t_1)}(t_1) + \delta \right\}. \tag{3.1}$$

5. **Baseline strategy update:** Each player updates a baseline action as follows:
   - If $\mathcal{A}_i^*(t_1) = \emptyset$, then $a_i^b(t_1+1) = a_i^b(t_1)$.
   - If $\mathcal{A}_i^*(t_1) \neq \emptyset$, then
     - with probability ω, set $a_i^b(t_1+1) = a_i^b(t_1)$. (We will refer to ω as the player's inertia.)
     - with probability $1-\omega$, randomly select $a_i^b(t_1+1) \in \mathcal{A}_i^*(t_1)$ with uniform probability.
6. Return to step 2 and repeat.

For simplicity, we will first state and prove the desired convergence properties using noiseless utility measurements. The setup for the noisy utility measurements will be stated afterwards.

Before stating the following theorem, we define the constant $\alpha > 0$ as follows. If $U_i(a^1) \neq U_i(a^2)$ for any joint actions $a^1, a^2 \in \mathcal{A}$ and any player $\mathcal{P}_i \in \mathcal{P}$, then $|U_i(a^1) - U_i(a^2)| > \alpha$. In other words, if any two joint actions result in different utilities at all, then the difference would be at least α.

Theorem 3.3. *Let G be a finite n-player weakly acyclic game in which all players use the sample experimentation dynamics. For any*

- *probability $p < 1$,*
- *tolerance level $\delta \in (0, \alpha)$,*
- *inertia $\omega \in (0,1)$, and*
- *exploration rate ϵ satisfying $\min\{(\alpha-\delta)/4, \delta/4, 1-p\} > (1-(1-\epsilon)^n) > 0$,*

if the exploration phase length m is sufficiently large, then for all sufficiently large times $t > 0$, $a(t)$ is a Nash equilibrium of G with at least probability p.

The remainder of this subsection is devoted to the proof of Theorem 3.3.

We will assume for simplicity that utilities are between $-1/2$ and $1/2$, i.e., $|U_i(a)| \leq 1/2$ for any player $\mathcal{P}_i \in \mathcal{P}$ and any joint action $a \in \mathcal{A}$.

We begin with a series of useful claims. The first claim states that for any player $\mathcal{P}_i$ the average utility for an action $a_i \in \mathcal{A}_i$ during the exploration phase can be made arbitrarily close (with high probability) to the actual utility the player would have received provided that all other players never experimented. This can be accomplished if the experimentation rate is sufficiently small and the exploration phase length is sufficiently large.

Claim 3.4. *Let a^b be the joint baseline action at the start of an exploration phase of length m. For*

- *any probability $p < 1$,*
- *any $\delta^* > 0$, and*
- *any exploration rate $\epsilon > 0$ satisfying $\delta^*/2 \geq (1-(1-\epsilon)^{n-1}) > 0$,*

if the exploration phase length m is sufficiently large, then

$$\mathbf{Pr}\left[\left|\hat{V}_i^{a_i} - U_i(a_i, a_{-i}^b)\right| > \delta^*\right] < 1 - p.$$

Proof. Let $n_i(a_i)$ represent the number of times player $\mathcal{P}_i$ played action a_i during the exploration phase. In the following discussion, *all probabilities and expectations are conditioned on* $n_i(a_i) > 0$. We omit making this explicit for the sake of notational simplicity. The event $n_i(a_i) = 0$ has diminishing probability as the exploration phase length m increases, and so this case will not affect the desired conclusions for increasing phase lengths.

For an arbitrary $\delta^* > 0$,

$$\begin{aligned}
&\mathbf{Pr}\left[\left|\hat{V}_i^{a_i} - U_i(a_i, a_{-i}^b)\right| > \delta^*\right] \\
&\quad \leq \mathbf{Pr}\left[\left|\hat{V}_i^{a_i} - E\{\hat{V}_i^{a_i}\}\right| + \left|E\{\hat{V}_i^{a_i}\} - U_i(a_i, a_{-i}^b)\right| > \delta^*\right] \\
&\quad \leq \underbrace{\mathbf{Pr}\left[\left|\hat{V}_i^{a_i} - E\{\hat{V}_i^{a_i}\}\right| > \delta^*/2\right]}_{(*)} + \underbrace{\mathbf{Pr}\left[\left|E\{\hat{V}_i^{a_i}\} - U_i(a_i, a_{-i}^b)\right| > \delta^*/2\right]}_{(**)}.
\end{aligned}$$

First, let us focus on $(**)$. We have

$$E\{\hat{V}_i^{a_i}\} - U_i(a_i, a_{-i}^b) = [1 - (1-\epsilon)^{n-1}]\left[E\{U_i(a_i, a_{-i}(t)) \mid a_{-i}(t) \neq a_{-i}^b\} - U_i(a_i, a^b)\right],$$

which approaches 0 as $\epsilon \downarrow 0$. Therefore, for any exploration rate ϵ satisfying $\delta^*/2 > (1 - (1-\epsilon)^{n-1}) > 0$, we know that

$$\mathbf{Pr}\left[\left|E\{\hat{V}_i^{a_i}\} - U_i(a_i, a_{-i}^b)\right| > \delta^*/2\right] = 0.$$

Now we will focus on $(*)$. By the weak law of large numbers, $(*)$ approaches 0 as $n_i(a_i) \uparrow \infty$. This implies that for any probability $\bar{p} < 1$ and any exploration rate $\epsilon > 0$, there exists a sample size $n_i^*(a_i)$ such that if $n_i(a_i) > n_i^*(a_i)$, then

$$\mathbf{Pr}\left[\left|\hat{V}_i^{a_i} - E\{\hat{V}_i^{a_i}\}\right| > \rho/2\right] < 1 - \bar{p}.$$

Lastly, for any probability $\bar{p} < 1$ and any fixed exploration rate, there exists a minimum exploration length $\underline{m} > 0$ such that for any exploration length $m > \underline{m}$,

$$\mathbf{Pr}\left[n_i(a_i) \geq n_i^*(a_i)\right] \geq \bar{p}.$$

In summary, for any fixed exploration rate ϵ satisfying $\delta^*/2 \geq (1 - (1-\epsilon)^{n-1}) > 0$, $(*) + (**)$ can be made arbitrarily close to 0, provided that the exploration length m is sufficiently large. ◻

CLAIM 3.5. *Let a^b be the joint baseline action at the start of an exploration phase of length m. For any*

- *probability $p < 1$,*
- *tolerance level $\delta \in (0, \alpha)$, and*
- *exploration rate $\epsilon > 0$ satisfying* $\min\{(\alpha - \delta)/4, \delta/4\} \geq (1 - (1-\epsilon)^{n-1}) > 0$,

if the exploration length m is sufficiently large, then each player's better response set $\mathcal{A}_i^*$ will contain only and all *actions that are a better response to the joint baseline action, i.e.,*

$$a_i^* \in \mathcal{A}_i^* \Leftrightarrow U_i(a_i^*, a_{-i}^b) > U_i(a^b)$$

with at least probability p.

Proof. Suppose a^b is not a Nash equilibrium. For some player $\mathcal{P}_i \in \mathcal{P}$, let a_i^* be a strict better reply to the baseline joint action, i.e., $U_i(a_i^*, a_{-i}^b) > U_i(a^b)$, and let a_i^w be a nonbetter reply to the baseline joint action, i.e., $U_i(a_i^w, a_{-i}^b) \leq U_i(a^b)$.

Using Claim 3.4, for any probability $\bar{p} < 1$ and any exploration rate $\epsilon > 0$ satisfying $\min\{(\alpha - \delta)/4, \delta/4\} \geq (1 - (1 - \epsilon)^{n-1}) > 0$ there exists a minimum exploration length $\underline{m} > 0$ such that for any exploration length $m > \underline{m}$ the following expressions are true:

$$\mathbf{Pr}\left[|\hat{V}_i^{a_i^b} - U_i(a_i^b, a_{-i}^b)| < \delta^*\right] \geq \bar{p}, \tag{3.2}$$

$$\mathbf{Pr}\left[|\hat{V}_i^{a_i^*} - U_i(a_i^*, a_{-i}^b)| < \delta^*\right] \geq \bar{p}, \tag{3.3}$$

$$\mathbf{Pr}\left[|\hat{V}_i^{a_i^w} - U_i(a_i^w, a_{-i}^b)| < \delta^*\right] \geq \bar{p}, \tag{3.4}$$

where $\delta^* = \min\{(\alpha - \delta)/2, \delta/2\}$. Rewriting (3.2), we obtain

$$\mathbf{Pr}\left[|\hat{V}_i^{a_i^b} - U_i(a_i^b, a_{-i}^b)| < \delta^*\right] \leq \mathbf{Pr}\left[\hat{V}_i^{a_i^b} - U_i(a_i^b, a_{-i}^b) < (\alpha - \delta)/2\right],$$

and rewriting (3.3), we obtain

$$\begin{aligned}\mathbf{Pr}\left[|\hat{V}_i^{a_i^*} - U_i(a_i^*, a_{-i}^b)| < \delta^*\right] &\leq \mathbf{Pr}\left[\hat{V}_i^{a_i^*} - U_i(a_i^*, a_{-i}^b) > -(\alpha - \delta)/2\right] \\ &\leq \mathbf{Pr}\left[\hat{V}_i^{a_i^*} - (U_i(a_i^b, a_{-i}^b) + \alpha) > -(\alpha - \delta)/2\right] \\ &= \mathbf{Pr}\left[\hat{V}_i^{a_i^*} - U_i(a_i^b, a_{-i}^b) > (\alpha + \delta)/2\right],\end{aligned}$$

meaning that

$$\mathbf{Pr}\left[a_i^* \in \mathcal{A}_i^*\right] \geq \bar{p}^2.$$

Similarly, rewriting (3.2), we obtain

$$\mathbf{Pr}\left[|\hat{V}_i^{a_i^b} - U_i(a_i^b, a_{-i}^b)| < \delta^*\right] \leq \mathbf{Pr}\left[\hat{V}_i^{a_i^b} - U_i(a_i^b, a_{-i}^b) > -\delta/2\right],$$

and rewriting (3.4), we obtain

$$\begin{aligned}\mathbf{Pr}\left[|\hat{V}_i^{a_i^w} - U_i(a_i^w, a_{-i}^b)| < \delta^*\right] &\leq \mathbf{Pr}\left[\hat{V}_i^{a_i^w} - U_i(a_i^w, a_{-i}^b) < \delta/2\right] \\ &\leq \mathbf{Pr}\left[\hat{V}_i^{a_i^w} - U_i(a_i^b, a_{-i}^b) < \delta/2\right],\end{aligned}$$

meaning that

$$\mathbf{Pr}\left[a_i^w \notin \mathcal{A}_i^*\right] \geq \bar{p}^2.$$

Since $\bar{p}$ can be chosen arbitrarily close to 1, the proof is complete. $\square$

Proof of Theorem 3.3. The evolution of the baseline actions from phase to phase is a finite aperiodic Markov process on the state space of joint actions, $\mathcal{A}$. Furthermore, since G is weakly acyclic, from every state there exists a better reply path to a Nash equilibrium. Hence, every recurrent class has at least one Nash equilibrium. We will show that these dynamics can be viewed as a perturbation of a certain Markov chain whose recurrent classes are restricted to Nash equilibria. We will then appeal to Theorem A.1 to derive the desired result.

We begin by defining an "unperturbed" process on baseline actions. For any $a^b \in \mathcal{A}$, define the *true* better reply set as

$$\bar{\mathcal{A}}_i^*(a^b) = \left\{a_i : U_i(a_i, a_{-i}^b) > U_i(a^b)\right\}.$$

Now define the transition process from $a^b(t_1)$ to $a^b(t_1+1)$ as follows:

- If $\bar{\mathcal{A}}_i^*(a^b(t_1)) = \emptyset$, then $a_i^b(t_1+1) = a_i^b(t_1)$.
- If $\bar{\mathcal{A}}_i^*(a^b(t_1)) \neq \emptyset$, then
  - with probability ω, set $a_i^b(t_1+1) = a_i^b(t_1)$.
  - with probability $1-\omega$, randomly select $a_i^b(t_1+1) \in \bar{\mathcal{A}}_i^*(t_1)$ with uniform probability.

This is a special case of a so-called "better reply process with finite memory and inertia." From [26, Theorem 6.2], the joint actions of this process converge to a Nash equilibrium with probability 1 in any weakly acyclic game. Therefore, the recurrence classes of this unperturbed are precisely the set of pure Nash equilibria.

The above unperturbed process closely resembles the baseline strategy update process described in step 5 of sample experimentation dynamics. The difference is that the above process uses the true better reply set, whereas step 5 uses a better reply set constructed from experimentation over a phase. However, by Claim 3.5, for any probability $\bar{p} < 1$, acceptable tolerance level δ, and acceptable exploration rate ϵ, there exists a minimum exploration phase length $\underline{m}$ such that for any exploration phase length $m > \underline{m}$, each player's better response set will contain only and all actions that are a strict better response with at least probability $\bar{p}$.

With parameters selected according to Claim 3.5, the transitions of the baseline joint actions in sample experimentation dynamics follow that of the above unperturbed better reply process with probability $\bar{p}$ arbitrarily close to 1. Since the recurrence classes of the unperturbed process are only Nash equilibria, we can conclude from Theorem A.1 that as $\bar{p}$ approaches 1, the probability that the baseline action for sufficiently large t_1 will be a (pure) Nash equilibrium can be made arbitrarily close to 1. By selecting the exploration probability ϵ sufficiently small, we can also conclude that the joint action during exploration phases, i.e., $a(mt_1 + t_2)$, will also be a Nash equilibrium with probability arbitrarily close to 1. □

3.3.2. Noisy utility measurements. Suppose that each player receives a noisy measurement of his true utility, i.e.,

$$\tilde{U}_i(a_i, a_{-i}) = U_i(a_i, a_{-i}) + \nu_i,$$

where n_i is an i.i.d. random variable with zero mean. In the regret testing algorithm with noisy utility measurements, the average utility for action a_i during the exploration phase at time t_1 is now

$$\hat{V}_i^{a_i}(t_1) = \begin{cases} \frac{1}{n_i^{a_i}(t_1)} \sum_{t_2=1}^{m} I\{a_i = \check{a}_i(t_1, t_2)\}\tilde{U}_i(\check{a}(t_1, t_2)), & n_i^{a_i}(t_1) > 0, \\ U_{\min}, & n_i^{a_i}(t_1) = 0. \end{cases}$$

A straightforward modification of the proof of Theorem 3.3 leads to the following theorem.

THEOREM 3.4. *Let G be a finite n-player weakly acyclic game where players' utilities are corrupted with a zero mean noise process. If all players use the sample experimentation dynamics, then for any*

- *probability* $p < 1$,
- *tolerance level* $\delta \in (0, \alpha)$,
- *inertia* $\omega \in (0, 1)$, *and*
- *exploration rate* ϵ *satisfying* $\min\{(\alpha - \delta)/4, \delta/4, 1 - p\} > (1 - (1 - \epsilon)^n) > 0$,

if the exploration phase length m *is sufficiently large, then for all sufficiently large times* $t > 0$, $a(t)$ *is a Nash equilibrium of* G *with at least probability* p.

3.3.3. Comment on length and synchronization of players' exploration phases. In the proof of Theorem 3.3, we assumed that all players' exploration phases were synchronized and of the same length. This assumption was used to ensure that when a player assessed the performance of a particular action, the baseline action of the other players remained constant. Because of the players' inertia this assumption is unnecessary. The general idea is as follows: a player will repeat a baseline action regardless of the better response set with positive probability because of the inertia. Therefore, if all players repeat their baseline action a sufficient number of times, which happens with positive probability, then the joint baseline action would remain constant long enough for any player to evaluate an accurate better response set for that particular joint baseline action.

4. Influencing Nash equilibria in resource allocation problems. In this section we will derive an approach for influencing the Nash equilibria of a resource allocation problem using the idea of marginal cost pricing. We will illustrate the setup and our approach on a congestion game which is an example of a resource allocation problem.

4.1. Congestion game setup. We consider a transportation network with a finite set R of road segments (or resources) that needs to be shared by a set of selfish drivers labeled as $D := \{d_1, \ldots, d_n\}$. Each driver has a fixed origin/destination pair connected through multiple routes. The set of all routes available to driver d_i is denoted by $\mathcal{A}_i$. A route $a_i \in \mathcal{A}_i$ consists of multiple road segments, therefore, $a_i \subset R$. Player $\mathcal{P}_i$ taking route a_i incurs a cost c_r for each road segment $r \in a_i$. The utility of driver d_i taking route a_i is defined as the negative of the total cost incurred, i.e., $U_i = -\sum_{r \in a_i} c_r$. Of course, the utility of each driver will depend on the routes chosen by other drivers.

If we assume that the cost incurred in a road segment depends *only* on the total number of drivers sharing that road, then drivers are anonymous, and this leads to a *congestion game* [19]. The utility of driver d_i is now stated more precisely as

$$U_i(a) = -\sum_{r \in a_i} c_r(\sigma_r(a)),$$

where $a := (a_1, \ldots, a_n)$ is the profile of routes chosen by all drivers and $\sigma_r(a)$ is the total number of drivers using the road segment r.

It is known that a congestion game admits the following potential function:

$$\hat{\phi}(a) = \sum_{r \in R} \sum_{k=1}^{\sigma_r(a)} c_r(k).$$

Unfortunately, this potential function lacks practical significance for measuring the effectiveness of a routing strategy in terms of the overall congestion.

4.2. Congestion game with tolls setup. One approach for equilibrium manipulation is to influence drivers' utilities with tolls [21]. In a congestion game with tolls, a driver's utility takes on the form

$$U_i(a) = -\sum_{r \in a_i} c_r(\sigma_r(a)) + t_r(\sigma_r(a)),$$

where $t_r(k)$ is the toll imposed on route r if there are k users.

Suppose that the global planner is interested in minimizing the total congestion experienced by all drivers on the network, which can be evaluated as

$$T_c(a) := \sum_{r \in R} \sigma_r(a) c_r(\sigma_r(a)).$$

It has been shown that there exists a set of tolls such that the potential function associated with the congestion game with tolls is aligned with the total congestion experienced by all drivers on the network (see [15, Proposition 4.1]).

PROPOSITION 4.1. *Consider a congestion game of any network topology. If the imposed tolls are set as*

$$t_r(k) = (k-1)[c_r(k) - c_r(k-1)] \quad \forall k \geq 1,$$

then the total negative congestion experienced by all drivers, $\phi_c(a) = -T_c(a)$, is a potential function for the congestion game with tolls.

This tolling scheme results in drivers' local utility functions being aligned with the global objective of minimal total congestion.

Now suppose that the global planner is interested in minimizing a *more general* measure,[4]

$$\phi(a) := \sum_{r \in R} f_r(\sigma_r(a)) c_r(\sigma_r(a)), \tag{4.1}$$

where $f_r : \{0, 1, 2, \ldots\} \to \mathbb{R}$ is any arbitrary function. An example of an objective function that fits within this framework and may be practical for general resource allocation problems is

$$\phi(a) = \sum_{r \in R} c_r(\sigma_r(a)).$$

We will now show that there exists a set of tolls, $t_r(\cdot)$, such that the potential function associated with the congestion game with tolls will be aligned with the global planner's objective function of the form given in (4.1).

PROPOSITION 4.2. *Consider a congestion game of any network topology. If the imposed tolls are set as*

$$t_r(k) = (f_r(k) - 1) c_r(k) - f_r(k-1) c_r(k-1) \quad \forall k \geq 1,$$

then the global planners objective, $\phi_c(a) = -\phi(a)$, is a potential function for the congestion game with tolls.

[4] In fact, if $c_r(\sigma_r(a)) \neq 0$ for all a, then (4.1) is equivalent to $\sum_{r \in R} f_r(\sigma_r(a))$.

Proof. Let $a^1 = \{a_i^1, a_{-i}\}$ and $a^2 = \{a_i^2, a_{-i}\}$. We will use the shorthand notation $\sigma_r^{a^1}$ to represent $\sigma_r(a^1)$. The change in utility incurred by driver d_i in changing from route a_i^2 to route a_i^1 is

$$\begin{aligned}U_i(a^1) - U_i(a^2) &= -\sum_{r \in a_i^1} \big(c_r(\sigma_r^{a^1}) + t_r(\sigma_r^{a^1})\big) + \sum_{r \in a_i^2} \big(c_r(\sigma_r^{a^2}) + t_r(\sigma_r^{a^2})\big) \\ &= -\sum_{r \in a_i^1 \backslash a_i^2} \big(c_r(\sigma_r^{a^1}) + t_r(\sigma_r^{a^1})\big) + \sum_{r \in a_i^2 \backslash a_i^1} \big(c_r(\sigma_r^{a^2}) + t_r(\sigma_r^{a^2})\big).\end{aligned}$$

The change in the total negative congestion from the joint action a^2 to a^1 is

$$\phi_c(a^1) - \phi_c(a^2) = -\sum_{r \in (a_i^1 \cup a_i^2)} \big(f_r(\sigma_r^{a^1}) c_r(\sigma_r^{a^1}) - f_r(\sigma_r^{a^2}) c_r(\sigma_r^{a^2})\big).$$

Since

$$\sum_{r \in (a_i^1 \cap a_i^2)} \big(f_r(\sigma_r^{a^1}) c_r(\sigma_r^{a^1}) - f_r(\sigma_r^{a^2}) c_r(\sigma_r^{a^2})\big) = 0,$$

the change in the total negative congestion is

$$\begin{aligned}&\phi_c(a^1) - \phi_c(a^2) \\ &\quad = -\sum_{r \in a_i^1 \backslash a_i^2} \big(f_r(\sigma_r^{a^1}) c_r(\sigma_r^{a^1}) - f_r(\sigma_r^{a^2}) c_r(\sigma_r^{a^2})\big) \\ &\qquad - \sum_{r \in a_i^2 \backslash a_i^1} \big(f_r(\sigma_r^{a^1}) c_r(\sigma_r^{a^1}) - f_r(\sigma_r^{a^2}) c_r(\sigma_r^{a^2})\big).\end{aligned}$$

Expanding the first term, we obtain

$$\begin{aligned}&\sum_{r \in a_i^1 \backslash a_i^2} \big(f_r(\sigma_r^{a^1}) c_r(\sigma_r^{a^1}) - f_r(\sigma_r^{a^2}) c_r(\sigma_r^{a^2})\big) \\ &\qquad = \sum_{r \in a_i^1 \backslash a_i^2} \big(f_r(\sigma_r^{a^1}) c_r(\sigma_r^{a^1}) - (f_r(\sigma_r^{a^1} - 1)) c_r(\sigma_r^{a^1} - 1)\big) \\ &\qquad = \sum_{r \in a_i^1 \backslash a_i^2} \big(f_r(\sigma_r^{a^1}) c_r(\sigma_r^{a^1}) - ((f_r(\sigma_r^{a^1}) - 1) c_r(\sigma_r^{a^1}) - t_r(\sigma_r^{a^1}))\big) \\ &\qquad = \sum_{r \in a_i^1 \backslash a_i^2} \big(c_r(\sigma_r^{a^1}) + t_r(\sigma_r^{a^1})\big).\end{aligned}$$

Therefore,

$$\begin{aligned}\phi_c(a^1) - \phi_c(a^2) &= -\sum_{r \in a_i^1 \backslash a_i^2} \big(c_r(\sigma_r^{a^1}) + t_r(\sigma_r^{a^1})\big) + \sum_{r \in a_i^2 \backslash a_i^1} \big(c_r(\sigma_r^{a^2}) + t_r(\sigma_r^{a^2})\big) \\ &= U_i(a^1) - U_i(a^2). \qquad \square\end{aligned}$$

By implementing the tolling scheme set forth in Proposition 4.2, we guarantee that all action profiles that minimize the global planner's objective are equilibrium of the congestion game with tolls.

In the special case that $f_r(\sigma_r(a)) = \sigma_r(a)$, Proposition 4.2 produces the same tolls as Proposition 4.1.

5. Illustrative example—congestion game.

We will consider a discrete representation of the congestion game setup considered in Braess' paradox [3]. In our setting, there are 1000 vehicles that need to traverse through the network. The network topology and associated congestion functions are illustrated in Figure 6. Each vehicle can select one of the four possible paths to traverse across the network.

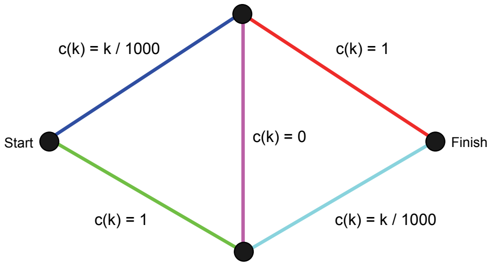


FIG. 6. *Congestion game setup.*

The reason for using this setup as an illustration of the learning algorithms and equilibrium manipulation approach developed in this paper is that the Nash equilibrium of this particular congestion game is easily identifiable. The unique Nash equilibrium is when all vehicles take the route as highlighted in Figure 7. At this Nash equilibrium each vehicle has a utility of 2 and the total congestion is 2000.

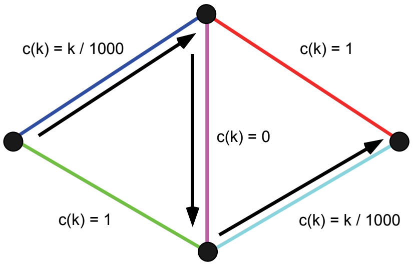


FIG. 7. *Illustration of Nash equilibrium in proposed congestion game.*

Since a potential game is weakly acyclic, the payoff-based learning dynamics in this paper are applicable learning algorithms for this congestion game. In a congestion game, a payoff-based learning algorithm means that drivers have access *only* to the actual congestion experienced. Drivers are unaware of the congestion level on any alternative routes. Figure 8 shows the evolution of drivers on routes when using the simple experimentation dynamics. This simulation used an experimentation rate of $\epsilon = 0.25\%$. One can observe that the vehicles' collective behavior does indeed approach that of the Nash equilibrium.

In this congestion game, it is also easy to verify that this vehicle distribution does not minimize the total congestion experience by all drivers over the network. The distribution that minimizes the total congestion over the network is when half the

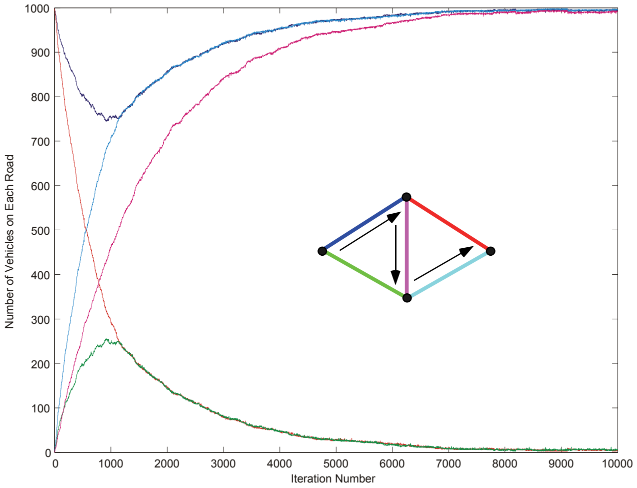


FIG. 8. *Evolution of number of vehicles on each road using simple experimentation dynamics: the number of vehicles on the roads highlighted by arrows approaches* 1000 *while the number of vehicles on all remaining roads approaches* 0.

vehicles occupy the top two roads and the other half occupy the bottom two roads. The middle road is irrelevant.

One can employ the tolling scheme developed in the previous section to locally influence vehicle behavior to achieve this objective. In this setting, the new cost functions, i.e., congestion plus tolls, are illustrated in Figure 9.

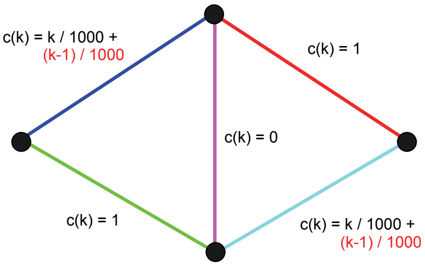


FIG. 9. *Congestion game setup with tolls to minimize total congestion.*

Figure 10 shows the evolution of drivers on routes when using the simple experimentation dynamics. This simulation used an experimentation rate of $\epsilon = 0.25\%$. When using this tolling scheme, the vehicles' collective behavior approaches the new Nash equilibrium which now minimizes the total congestion experienced on the network. The total congestion experienced on the network is now approximately 1500.

There are other tolling schemes that would have resulted in the desired allocation. One approach is to assign an infinite cost to the middle road, which is equivalent to removing it from the network. Under this scenario, the unique Nash equilibrium is for half the vehicles to occupy the top route and the other half to occupy the bottom,

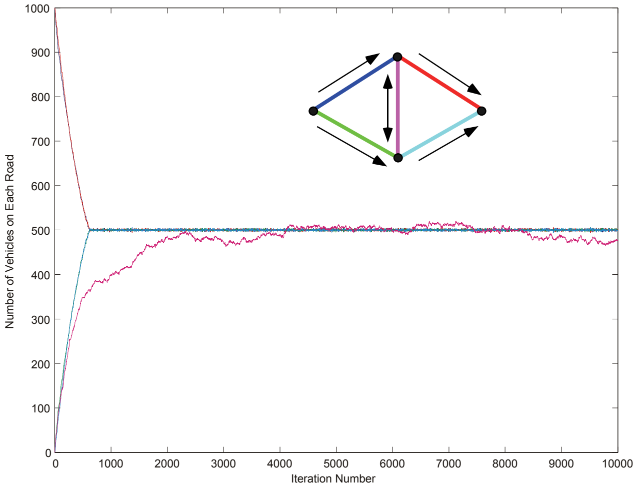


FIG. 10. *Evolution of number of vehicles on each road using simple experimentation dynamics with optimal tolls: the number of vehicles on the middle road fluctuates around* 500 *while the number of vehicles on all remaining roads stabilizes to around* 500*.*

which would minimize the total congestion on the network. Therefore, the existence of this extra road, even though it has zero cost, resulted in the unique Nash equilibrium having a higher total congestion. This is Braess' paradox [3].

The advantage of the tolling scheme set forth in this paper is that it gives a systematic method for influencing the Nash equilibria of any congestion game. We would like to highlight that this tolling scheme guarantees only that the action profiles that maximize the desired objective function are Nash equilibria of the new congestion game with tolls. However, it does not guarantee the lack of suboptimal Nash equilibria.

In many applications, players may not have access to their true utility, but do have access to a noisy measurement of their utility. For example, in the traffic setting, this noisy measurement could be the result of accidents or weather conditions. We will revisit the original congestion game (without tolls) as illustrated in Figure 6. We will now assume that a driver's utility measurement takes on the form

$$\tilde{U}_i(a) = -\sum_{r \in a_i} c_r(\sigma_r(a)) + \nu_i,$$

where ν_i is a random variable with zero mean and variance of 0.1. We will assume that the noise is driver specific rather than road specific.

Figure 11 shows a comparison of the evolution of drivers on routes when using the simple and sample experimentation dynamics. The simple experimentation dynamics simulation used an experimentation rate $\epsilon = 0.25\%$. The sample experimentation dynamics simulation used an exploration rate $\epsilon = 0.25\%$, a tolerance level $\delta = 0.002$, an exploration phase length $m = 500000$, and inertia $\omega = 0.85$. As expected, the noisy utility measurements influenced vehicle behavior more in the simple experimentation dynamics than the sample experimentation dynamics.

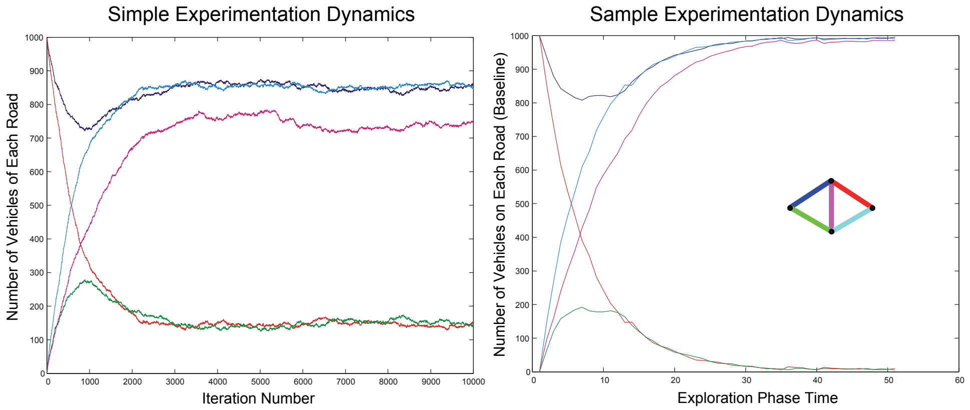


FIG. 11. *Comparison of evolution of number of vehicles on each road using simple experimentation dynamics and sample experimentation dynamics (baseline) with noisy utility measurements: the number of vehicles on the route (upper left, middle, lower right) dominates the number of vehicles on all remaining roads in both settings.*

6. Concluding remarks. We have introduced safe experimentation dynamics for identical interest games, simple experimentation dynamics for weakly acyclic games with noise-free utility measurements, and sample experimentation dynamics for weakly acyclic games with noisy utility measurements. For all three settings, we have shown that for sufficiently large times, the joint action taken by all players will constitute a Nash equilibrium. Furthermore, we have shown how to guarantee that a collective objective in a congestion game is a (nonunique) Nash equilibrium. An important, but unaddressed, topic in this work is characterizing resulting convergence rates. It is likely that tools regarding mixing times of Markov chains [17] will be relevant.

Our motivation has been that in many engineered systems, the functional forms of utility functions are not available, and so players must adjust their strategies through an adaptive process using only payoff measurements. In the dynamic processes defined here, there is no explicit cooperation or communication between players. On the one hand, this lack of explicit coordination offers an element of robustness to a variety of uncertainties in the strategy adjustment processes. Nonetheless, on the other hand, an interesting future direction would be to investigate to what degree explicit coordination through limited communications could be beneficial.

Appendix. Background on resistance trees. For a detailed review of the theory of resistance trees, please see [24].

Let P^0 denote the probability transition matrix for a finite state Markov chain over the state space Z. Consider a "perturbed" process such that the size of the perturbations can be indexed by a scalar $\epsilon > 0$, and let P^ϵ be the associated transition probability matrix. The process P^ϵ is called a *regular perturbed Markov process* if P^ϵ is ergodic for all sufficiently small $\epsilon > 0$ and P^ϵ approaches P^0 at an exponentially smooth rate [24]. Specifically, the latter condition means that for all $z, z' \in Z$,

$$\lim_{\epsilon \to 0^+} P^\epsilon_{zz'} = P^0_{zz'},$$

and

$$P^\epsilon_{zz'} > 0 \text{ for some } \epsilon > 0 \;\Rightarrow\; 0 < \lim_{\epsilon \to 0^+} \frac{P^\epsilon_{zz'}}{\epsilon^{r(z \to z')}} < \infty$$

for some nonnegative real number $r(z \to z')$, which is called the *resistance* of the transition $z \to z'$. (Note in particular that if $P^0_{zz'} > 0$, then $r(z \to z') = 0$.)

Let the recurrence classes of P^0 be denoted by $E_1, E_2, \ldots, E_N$. For each pair of distinct recurrence classes E_i and E_j, $i \neq j$, an ij-path is defined to be a sequence of distinct states $\zeta = (z_1 \to z_2 \to \cdots \to z_n)$ such that $z_1 \in E_i$ and $z_n \in E_j$. The resistance of this path is the sum of the resistances of its edges, that is, $r(\zeta) = r(z_1 \to z_2) + r(z_2 \to z_3) + \cdots + r(z_{n-1} \to z_n)$. Let $\rho_{ij} = \min r(\zeta)$ be the least resistance over all ij-paths ζ. Note that ρ_{ij} must be positive for all distinct i and j, because there exists no path of zero resistance between distinct recurrence classes.

Now construct a complete directed graph with N vertices, one for each recurrence class. The vertex corresponding to class E_j will be called j. The weight on the directed edge $i \to j$ is ρ_{ij}. A tree, T, rooted at vertex j, also called a j-tree, is a set of $N-1$ directed edges such that, from every vertex different from j, there is a unique directed path in the tree to j. The resistance of a rooted tree, T, is the sum of the resistances ρ_{ij} on the $N-1$ edges that compose it. The *stochastic potential*, γ_j, of the recurrence class E_j is defined to be the minimum resistance over all trees rooted at j. The following theorem gives a simple criterion for determining the stochastically stable states (see [24, Theorem 4]).

THEOREM A.1. *Let P^ϵ be a regular perturbed Markov process, and for each $\epsilon > 0$ let μ^ϵ be the unique stationary distribution of P^ϵ. Then $\lim_{\epsilon \to 0} \mu^\epsilon$ exists and the limiting distribution μ^0 is a stationary distribution of P^0. The stochastically stable states (i.e., the support of μ^0) are precisely those states contained in the recurrence classes with minimum stochastic potential.*

REFERENCES

[1] G. Arslan, J. R. Marden, and J. S. Shamma, *Autonomous vehicle-target assignment: A game theoretical formulation*, ASME J. Dynam. Systems Measurement and Control, 129 (2007), pp. 584–596.

[2] V. S. Borkar and P. R. Kumar, *Dynamic Cesaro-Wardrop equilibration in networks*, IEEE Trans. Automat. Control, 48 (2003), pp. 382–396.

[3] D. Braess, *Uber ein paradoxen der verkehrsplanning*, Unternehmensforschung, 12 (1968), pp. 258–268.

[4] D. P. Foster and H. P. Young, *Regret testing: Learning to play Nash equilibrium without knowing you have an opponent*, Theoret. Econom., 1 (2006), pp. 341–367.

[5] D. Fudenberg and D. K. Levine, *The Theory of Learning in Games*, MIT Press, Cambridge, MA, 1998.

[6] D. Fudenberg and J. Tirole, *Game Theory*, MIT Press, Cambridge, MA, 1991.

[7] A. Ganguli, S. Susca, S. Martinez, F. Bullo, and J. Cortes, *On collective motion in sensor networks: Sample problems and distributed algorithms*, in Proceedings of the 44th IEEE Conference on Decision and Control, Seville, Spain, 2005, pp. 4239–4244.

[8] F. Germano and G. Lugosi, *Global Nash convergence of Foster and Young's regret testing*, Games Econom. Behavior, 60 (2007), pp. 135–154.

[9] S. B. Gershwin, *Manufacturing Systems Engineering*, Prentice-Hall, Englewood Cliffs, NJ, 1994.

[10] S. Hart, *Adaptive heuristics*, Econometrica, 73 (2005), pp. 1401–1430.

[11] J. Hofbauer and K. Sigmund, *Evolutionary Games and Population Dynamics*, Cambridge University Press, Cambridge, UK, 1998.

[12] S. Mannor and J. S. Shamma, *Multi-agent learning for engineers*, Artificial Intelligence, 171 (2007), pp. 417–422.

[13] J. R. Marden, G. Arslan, and J. S. Shamma, *Connections between cooperative control and potential games illustrated on the consensus problem*, in Proceedings of the 2007 European Control Conference (ECC '07), Kos, Greece, 2007, pp. 4604–4611.

[14] J. R. Marden, G. Arslan, and J. S. Shamma, *Regret based dynamics: Convergence in weakly acyclic games*, in Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS), Honolulu, HI, 2007, article 42.

[15] J. R. Marden, G. Arslan, and J. S. Shamma, *Joint strategy fictitious play with inertia for potential games*, IEEE Trans. Automat. Control, to appear.

[16] D. Monderer and L. S. Shapley, *Fictitious play property for games with identical interests*, J. Econom. Theory, 68 (1996), pp. 258–265.

[17] R. Montenegro and P. Tetali, *Mathematical Aspects of Mixing Times in Markov Chains*, Now Publishers, Hanover, MA, 2006.

[18] L. Panait and S. Luke, *Cooperative multi-agent learning: The state of the art*, Autonomous Agents and Multi-Agent Systems, 11 (2005), pp. 387–434.

[19] R. W. Rosenthal, *A class of games possessing pure-strategy Nash equilibria*, Internat. J. Game Theory, 2 (1973), pp. 65–67.

[20] L. Samuelson, *Evolutionary Games and Equilibrium Selection*, MIT Press, Cambridge, MA, 1997.

[21] W. Sandholm, *Evolutionary implementation and congestion pricing*, Rev. Econom. Stud., 69 (2002), pp. 667–689.

[22] Y. Shoham, R. Powers, and T. Grenager, *If multi-agent learning is the answer, what is the question?*, Artificial Intelligence, 171 (2007), pp. 365–377.

[23] J. W. Weibull, *Evolutionary Game Theory*, MIT Press, Cambridge, MA, 1995.

[24] H. P. Young, *The evolution of conventions*, Econometrica, 61 (1993), pp. 57–84.

[25] H. P. Young, *Individual Strategy and Social Structure*, Princeton University Press, Princeton, NJ, 1998.

[26] H. P. Young, *Strategic Learning and Its Limits*, Oxford University Press, Oxford, UK, 2005.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 1, pp. 397–413

© 2009 Society for Industrial and Applied Mathematics

MINIMAL INTERCONNECTION TOPOLOGY IN DISTRIBUTED CONTROL DESIGN[*]

CÉDRIC LANGBORT[†] AND VIJAY GUPTA[‡]

Abstract. In this paper, we consider a distributed control design problem. Multiple agents (or subsystems) that are dynamically uncoupled need to be controlled to optimize a joint cost function. An interconnection graph specifies the topology according to which the agents can access information about each others' state. We propose and partially analyze a new model for determining the influence of the topology of the interconnection graph on the performance achieved by the subsystems. We consider the classical linear-quadratic regulator (LQR) cost function and propose making one of the weight matrices to be topology dependent to capture the extra cost incurred when more communication between the agents is allowed. We present results about optimal topologies for some models of the dependence of the weight matrix on the communication graph. We also give some results about the existence of "critical prices" at which adding supplementary edges becomes detrimental to closed-loop performance. One conclusion of the work is that if the communication between the agents comes at a cost, then adding communication edges may be harmful for the system performance.

Key words. distributed control, interconnection topology

AMS subject classifications. 93A14, 93C41, 49N10

DOI. 10.1137/06067897X

1. Introduction.

The question of optimal decentralized or structured control design for systems composed of interconnected subsystems has been widely studied at least since the 1970s (see, e.g., [4, 20]). The defining feature of these problems is that, while optimization of the cost function can demand that the control of one subsystem know the states of all the other subsystems, the topology (or information pattern) imposed by the specified system structure may not allow such interactions to happen. A lot is known about the information patterns for which an optimal structured control law exists [24] or has (or does not have) some desirable properties, such as being linear and satisfying a separation principle [20, 25], or being computable in polynomial time in the dimension of the plant's data [5, 19]. Many synthesis methods are also available that can impose a specific topology on a controller (see, e.g., [15] and the references therein), resulting in control laws that can be proved to have a particular structure a posteriori [2, 8, 18] or that give an approximation to the exact optimal structured controller (see, e.g., [11] and the references therein).

In all the works mentioned above, however, the controller's interconnection topology is always (if sometimes implicitly) assumed to be known to the designer prior to synthesis, and optimization and/or design are to be performed among control laws with this particular structure. While this problem formulation is appropriate when decentralization is viewed as an external constraint (e.g., when controlling a system with a *pre-existing* interconnection topology such as a power distribution network), there are situations where the interconnection graph and the controller are both designed

[*] Received by the editors December 31, 2006; accepted for publication (in revised form) August 10, 2008; published electronically February 11, 2009. A preliminary version of these results was presented at the American Control Conference, 2006.

http://www.siam.org/journals/sicon/48-1/67897.html

[†] Department of Aerospace Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801 (langbort@illinois.edu).

[‡] Department of Electrical Engineering, University of Notre Dame, South Bend, IN 46556 (vgupta2@nd.edu).

at the same time. For example, when designing cooperative multi-agent systems, the choice of the architecture or the information pattern (leader-follower, fully decentralized, or some other structure) may itself be a valid design choice, and thus, an integral part of the control design process. In such cases, it makes sense to try to find the minimal (with respect to some cost function) topology needed to achieve a particular control goal.

This structural optimization problem is complementary to other communication theoretic trade-offs arising in network/distributed controller codesign (as discussed, e.g., in [16]). The optimization of the communication topology along with the control laws followed by the individual agents also falls within the framework of the theory of organizational efficiency and information cost introduced in economics by Marschak and Radner in [17] and later studied in, e.g., [6, 23, 14]. It is surprising that, while the tools and ideas of team decision theory that were developed in [17] have been successfully applied to distributed control problems with given information structure [13], the very question of efficiency of decision architectures, which these authors were originally interested in, has received relatively little attention in the control literature. In particular, we are only aware of the following recent attempts at finding a minimal control interconnection structure:

- In [21], the authors have shown that, when constructing a distributed controller from a set of observer-based controllers using different and parallel observations, the star interconnection topology is minimal, in the sense that the resulting control design problem has the minimal number of free parameters needed to ensure closed-loop stability.
- In [26], another kind of optimal topology design is investigated, which uses the closed-loop system's convergence rate to equilibrium as a cost function. More precisely, the authors use semidefinite programming to characterize the weighted interconnection graph resulting in the fastest convergence to consensus, under the assumption of nearest-neighbor averaging dynamics.

In this paper, we present a new model for studying the role of controller topology in distributed control problems, which is inspired by some of the works in the economics literature mentioned above.

More precisely, we propose to modify the classical linear-quadratic regulator (LQR) cost function by making one of the weight matrices topology dependent. While this way of explicitly incorporating the role of communication topologies in control design is admittedly somewhat artificial, it allows us to bring the tools of LQR design and convex optimization to bear on the problem and, in some cases, to give tractable algorithms for designing the minimal topology, even though the design of optimal structured controllers, *for a fixed topology*, is itself a difficult question.

Other possible topology or communication-dependent costs for the codesign of an information pattern and controller, which are not considered here, include

(i) a fixed one-shot cost for the addition of every new communication edge,
(ii) a price that is proportional to the bit or entropy rate of the messages transmitted along the edges.

In case (i), the global cost associated with the communication topology would be independent of the quantity of information exchanged between subcontrollers, while in case (ii), this quantity would be accounted for very precisely. The cost function considered in this paper thus represents an intermediate case, in which the price paid for communication *does* depend on the exchanged signals (although maybe in a nonphysical way), but where the details of the communication protocols used are not relevant.

The paper is organized as follows. We begin by introducing our model and various notations in section 2. As in most of the works on distributed control mentioned above, we assume the communication links if they exist, to be ideal. Then in section 3, we consider particular models of the dependence of the weight matrices on topology and compute the optimal cost achieved by various topologies. In particular, we prove the somewhat surprising result that, under certain assumptions, the optimal control topology according to our criterion is the fully decentralized one. Then, in section 4, we consider arbitrary graphs and arbitrary weight matrices. We provide some conditions for the existence of "critical prices" at which it becomes detrimental to add edges to a pre-existing controller topology. Finally, in section 5, we illustrate some of the results by considering a simple example.

2. Problem formulation.

2.1. Notation. For a matrix M, we will denote the (i,j)th element by $[M]_{ij}$. Similarly, for a column vector v, the ith element is denoted as v_i. For matrices and vectors that have been defined blockwise we will abuse the notation and use $[M]_{ij}$ to mean the (i,j)th block of M and use v_i to mean the ith block of v. The particular use will be clear from the context. The transpose of a matrix M and a vector v will be denoted by M^* and v^*, respectively. The norm of a vector v is denoted by the symbol $\|v\|$. For a matrix M, we denote the spectral radius of M by $\lambda_{\max}(M)$ and the maximum singular value by $\sigma_{\max}(M)$. Given matrices $M_1, M_2, \ldots, M_n$, we will denote the block-diagonal matrix, formed by placing the matrices M_i's along the diagonal, as $\mathbf{diag}_i\,(M_i)$. The space of all symmetric positive definite $n \times n$ matrices is denoted by $\mathbb{S}^n$. For two matrices M_1 and M_2, we will say $M_1 \geq M_2$ if $M_1 - M_2$ is positive semidefinite.

2.2. Structured control laws. Assume we are given N discrete time, linear time-invariant (LTI) subsystems (or agents) described by

$$x_i(k+1) = A_i x_i(k) + B_i u_i(k) \tag{2.1}$$

for all $i = 1 \ldots N$ and $k \geq 0$. At each time k, the state $x_i(k)$ and input $u_i(k)$ of each subsystem is an element of $\mathbb{R}^{n_i}$ and $\mathbb{R}^{m_i}$, respectively. In what follows, we will write $n := \sum_{i=1}^N n_i$ and $m := \sum_{i=1}^N m_i$. The state of the entire system can be defined by stacking the states of all the subsystems in a column vector, which we denote by $x(k)$:

$$x(k) = \begin{pmatrix} x_1(k) \\ \vdots \\ x_N(k) \end{pmatrix}.$$

We can similarly define the column vector $u(k)$ by stacking all the individual control inputs $u_i(k)$'s. Each pair (A_i, B_i) is assumed to be controllable which, in turn (see, e.g., [11]), implies that the full system, with matrices $A := \mathbf{diag}_i(A_i)$, $B := \mathbf{diag}_i(B_i)$, is also controllable.

Let $\mathcal{G}_N$ be the set of all undirected graphs with N vertices. We will think of each vertex of a graph $g \in \mathcal{G}_N$ as representing the subsystem (2.1) labeled with the same index i, where $1 \leq i \leq N$. To every graph $g \in \mathcal{G}_N$ is associated an edge set $E(g)$ defining the edges present in g and an adjacency matrix $\mathcal{A}(g)$ defined as

$$[\mathcal{A}(g)]_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E(g), \\ 0 & \text{otherwise.} \end{cases}$$

All the graphs we will consider have self-loops, i.e., $(i,i) \in E(g)$ for all $i = 1 \dots N$, $g \in \mathcal{G}_N$. If a graph g_1 is a subgraph of g_2, i.e., if $E(g_1) \subset E(g_2)$, we will write $g_1 \preccurlyeq g_2$. Clearly, relation " $\preccurlyeq$ " defines a partial order on $\mathcal{G}_N$.

Each graph in $\mathcal{G}_N$ specifies a communication topology that can be used to construct specific control laws for subsystems (2.1). To this end, we introduce the space $\mathcal{K}_{m,n}(g)$ of structured matrices with structure imposed by g. A matrix K in $\mathcal{K}_{m,n}(g)$ is defined blockwise, with each block K_{ij} being an $m_i \times n_j$ matrix such that $[K]_{ij} = 0$ whenever $[\mathcal{A}(g)]_{ij} = 0$. The space $\mathcal{K}_{n,n}(g)$ is defined similarly, with blocks of size $n_i \times n_j$.

According to these definitions, a control law u defined by

$$\begin{pmatrix} u_1(k) \\ \vdots \\ u_N(k) \end{pmatrix} = K(g) \begin{pmatrix} x_1(k) \\ \vdots \\ x_N(k) \end{pmatrix} \quad \text{for all } k \geq 0 \tag{2.2}$$

for some feedback matrix gain $K(g) \in \mathcal{K}_{m,n}(g)$ is such that $u_i(k)$ involves the values of $x_j(k)$ for *only* those j such that $(i,j) \in E(g)$.

Each edge (i,j) in a graph g can thus informally be thought of as a communication link allowing agents i and j to use each other's state value when computing their control input. The whole graph specifies the information pattern or the communication topology. A control law satisfying (2.2) is said to have *structure* g. An unstructured control law is one that has structure corresponding to the complete graph. It should also be noted from (2.2) that we are interested only in static and linear control laws so that the matrix $K(g)$ is time invariant. Thus, a communication link, if present, is assumed to be perfect in the sense that we ignore effects such as quantization issues, data dropouts, and data delays longer than one time step.

A control law of the form (2.2) will be called stabilizing if, in closed-loop form,

$$\lim_{k \to \infty} x(k) = 0 \quad \text{for all } x_0 \text{ such that } \|x_0\| \leq 1.$$

This is equivalent to the matrix $A + BK(g)$ being Schur, i.e., having all its eigenvalues with modulus strictly less than one.

2.3. Cost functions. For any given positive definite matrix $Q \in \mathbb{S}^n$, control law u, and initial condition x_0, we define the familiar quadratic cost

$$\begin{aligned} J(Q, x_0; u) &:= \frac{1}{2} \sum_{k=0}^{\infty} \begin{bmatrix} x_1(k) \\ \vdots \\ x_N(k) \end{bmatrix}^* Q \begin{bmatrix} x_1(k) \\ \vdots \\ x_N(k) \end{bmatrix} + \begin{bmatrix} u_1(k) \\ \vdots \\ u_N(k) \end{bmatrix}^* \begin{bmatrix} u_1(k) \\ \vdots \\ u_N(k) \end{bmatrix} \\ \text{subject to} \quad & x_i(k+1) = A_i x_i(k) + B_i u_i(k) \text{ for all } k \geq 0 \\ & x_i(0) = [x_0]_i \text{ for all } i = 1 \dots N. \end{aligned} \tag{2.3}$$

In order to study the influence of a control law's structure on this cost, it is preferable to eliminate the dependence on initial conditions and consider the worst-case cost

$$J(Q, u) := \sup_{\|x_0\| \leq 1} J(Q, x_0; u).$$

The minimum value of this worst-case cost over the choice of control laws is defined as

$$J^\star(Q) := \inf_u J(Q, u).$$

It is well known that for all initial conditions x_0 and control laws u, $J(Q, x_0; u) \geq x_0^* P x_0$, where P is the unique positive definite solution of the Riccati equation

$$P = A^*PA + Q - A^*PB\left(B^*PB + I\right)^{-1} B^*PA. \tag{2.4}$$

Furthermore, the cost $x_0^* P x_0$ is achievable by the familiar LQR control. Thus, we immediately obtain

$$J^\star(Q) = \lambda_{\max}(P).$$

When a graph $g \in \mathcal{G}_N$ imposes a structure on the allowed control laws, we can analogously define the best performance achievable by a structured control law as

$$J_g^\star(Q) := \begin{cases} \inf\limits_u & J(Q, u) \\ \text{subject to} & u \text{ has structure } g. \end{cases} \tag{2.5}$$

It is clear that for any graph g, $J^\star(Q) \leq J_g^\star(Q)$. In keeping with the spirit of the previous notations, we will write $J_g(Q, u)$ instead of $J(Q, u)$ when the control law u at hand has structure g. We will also sometimes abuse notation by writing $J_g(Q, K(g))$ instead of $J_g(Q, u)$ when the control law u and gain matrix $K(g)$ are related, as in (2.2).

We can now establish the following properties of function $J_g(Q, .)$ for a fixed positive definite matrix $Q \in \mathbb{S}^n$.

Proposition 2.1. *Let $g \in \mathcal{G}_N$ and matrix $Q \in \mathbb{S}^n$. A control law u with structure g is stabilizing if and only if $J_g(Q, u)$ is finite. Furthermore, when u is stabilizing, the cost $J_g(Q, u) = \lambda_{\max}(P(g))$, where the matrix $P(g)$ is the unique positive definite solution of the following Lyapunov equation:*

$$(A + BK(g))^* P(g)(A + BK(g)) - P(g) + Q + K(g)^* K(g) = 0, \tag{2.6}$$

with $K(g)$ being given as per (2.2) *for the control law u.*

Proof. If u is stabilizing, then Lyapunov equation (2.6) has a unique positive definite solution P for any matrix Q. Then, clearly, $J(Q, x_0; u) = x_0^* P x_0$ for all initial condition x_0 and, thus,

$$J_g(Q, u) = \sup_{\|x_0\| \leq 1} J_g(Q, x_0; u) = \lambda_{max}(P) < +\infty.$$

If u is not stabilizing, there exists an initial condition x_0 such that the state $x(k)$ of system (2.1) does not tend to zero as k goes to infinity in closed-loop form. Hence, $+\infty = J(Q, x_0; u) \leq J(Q, u)$. □

The following proposition shows that, under natural constraint qualification requirements, the definition of best performance achievable by a structured control law as defined in (2.5) makes sense.

Proposition 2.2. *Let $g \in \mathcal{G}_N$ be a graph and $Q \in \mathbb{S}^n$ a positive definite weight matrix. If there exists a stabilizing control law with structure g, then the infimum in* (2.5) *is attained; i.e., there exists an optimal control gain $\bar{K}(g) \in \mathcal{K}_{m,n}(g)$ such that $J_g^\star(Q) = J_g(Q, \bar{K}(g))$.*

Proof. Let $\{K_l\}$ be a minimizing sequence of control gains with structure g for $J_g(Q, .)$, i.e., such that

$$\lim_{l \to \infty} J_g(Q, K_l) = J_g^\star(Q).$$

Let u_0 be a stabilizing control law with structure g. By definition, there exists l_0 such that for $l > l_0$, $J_g(Q, K_l) \leq J_g(Q, u_0)$. Hence, for $l > l_0$ and all x_0 with norm less than one,

$$\begin{aligned} x_0^* K_l^* K_l x_0 &\leq J(Q, x_0; K_l) \\ &\leq J_g(Q, K_l) \\ &\leq J_g(Q, u_0) < +\infty, \end{aligned}$$

which means that, for $l > l_0$, K_l belongs to the closed ball (for the norm $\sigma_{max}(.)$)

$$\{K \in \mathcal{K}_{m,n}(g) \; | \sigma_{max}(K) \leq J_g(Q, u_0)\}.$$

Since $\mathcal{K}_{m,n}(g)$ is finite dimensional, this ball is compact, and thus $\{K_l\}$ has a converging subsequence (which we still denote as $\{K_l\}$) with limit $\bar{K}(g)$ in this ball. Finally, it is easy to show that function $J_g(Q, .)$ is continuous on its domain and, hence, that

$$J_g^\star(Q) = \lim_{n \to \infty} J_g(Q, K_n) = J_g(Q, \bar{K}(g)). \qquad \square$$

2.4. Value of a graph. We are now in a position to introduce the *value* of a graph. The terminology is borrowed from [14]. Let a mapping $\mathcal{Q} : \mathcal{G}_N \to \mathbb{S}^n$ be given. The value of graph g is defined as

$$V(g) := J_g^\star(\mathcal{Q}(g)). \tag{2.7}$$

The motivation for introducing such graph-dependent weighting matrices and cost functions is the following. Assume that we are interested in finding a controller minimizing the cost $J(Q_0, .)$ for some positive definite matrix Q_0. If there is no restriction on the structure of the control law, the optimal controller's interconnection topology will typically be a full graph. In practice, however, building and maintaining each of the graph's communication edges has a cost which, if taken into account, may make this control law less attractive. It is to capture this trade-off between closed-loop performance and controller topology that we introduce an *information cost* [17] associated to every communication graph g. Of course, there are many ways in which such a cost could be defined, and our choice of a graph-dependent weight matrix Q is mainly motivated by the fact that the resulting control design problem naturally fits into the LQR framework. Several choices are possible for the map $\mathcal{Q}$, some of which are detailed below along with a possible interpretation.

Definition 2.3. *We say that a weight map $\mathcal{Q} : \mathcal{G}_N \to \mathbb{S}^n$ is*

- edge separable without interference *if it satisfies*

$$\mathcal{Q}(g) := Q_0 + \sum_{\substack{(i,j) \in E(g) \\ i \neq j}} P_{ij} \tag{2.8}$$

for all graphs g, with each matrix $P_{ij} \geq 0$ being partitioned according to the subsystems and having all blocks, except the (i,i)th, (i,j)th, (j,i)th, and (j,j)th, equal to zero. In this case, a subsystem pays a price for transmitting the value of its state to another subsystem.

- edge separable with externalities *if it is still given by* (2.8)*, but with all the diagonal blocks of each matrix P_{ij} being possibly nonzero. Conceptually, such information costs can model situations in which the subsystems that are exchanging information over a particular link are not the only ones paying a* *price for it. Instead, the subsystems not directly involved in the communication also agree to contribute to the cost of information sharing. Such a model is reasonable when subsystems are cooperating with each other.*

The situation when the cost of one agent can be influenced by the action of others is referred to as an "externality" in the economics literature, which motivates the name used for this type of weight map.

- nonseparable or "with interference" *if $\mathcal{Q}(g)$ can be a full matrix without any particular structure for every graph g. This corresponds to the general case where the cost paid for communication over every link depends on all the other links present in the graph. It is to capture this idea of parasitic effect of edges on each other that we say that "there is interference." However, "interference" is not meant in its physical sense.*

Accounting for the communication cost through an increase in the value of the matrix Q implies that the communication cost varies as the square of the state value. As mentioned before, this quadratic communication cost is considered mostly for simplicity and ease of analysis. We note, however, that obtaining a measure of communication cost that is suitable for all applications of multi-agent systems and admits a theoretical analysis is still an open research question. We thus hope that our approach will provide a useful first step in this direction.

We will also make use of the following two properties of a map $\mathcal{Q}$.

DEFINITION 2.4. *We say that a map $\mathcal{Q} : \mathcal{G}_N \to \mathbb{S}^n$ is*

(i) nondecreasing *if for all $g, g' \in \mathcal{G}_N$, $g \preccurlyeq g' \quad \Rightarrow \quad \mathcal{Q}(g) \leq \mathcal{Q}(g')$.*

(ii) structure compatible *if for all $g \in \mathcal{G}_N$, $\mathcal{Q}(g) \in \mathcal{K}_{n,n}(g)$.*

Note that edge separability in the mapping $\mathcal{Q}$ is not sufficient for structure compatibility, since the matrix Q_0 in (2.8) may not be block diagonal.

Our goal in the following sections is to characterize the efficient graph $g^\star$ defined by

$$g^\star := \arg\min_{g \in \mathcal{G}_N} V(g). \tag{2.9}$$

The structure imposed by $g^\star$ corresponds to the minimal communication requirements (in the sense of the cost function (2.3)) needed to control the N agents. Since there are only finitely many elements in $\mathcal{G}_N$, $g^\star$ always exists provided the value function V is proper. One way to determine $g^\star$ would be to enumerate the value $V(g)$ for every graph $g \in \mathcal{G}_N$ and minimize over that set. The problems with this approach, however, are that calculating the value of a single graph is typically a hard, nonconvex optimization problem, and that the number of possible graphs in $\mathcal{G}_N$ is exponential in N.

In the remainder of this paper, we explore two alternative and complementary approaches for characterizing the efficient graph $g^\star$ which do not rely on such an exhaustive search. In section 3, we show that, for particular families of maps $\mathcal{Q}$, it is possible to compute $g^\star$ explicitly. Next, we consider more general maps $\mathcal{Q}$ and derive tractable necessary and sufficient conditions for two graphs g and g' to satisfy $V(g) \leq V(g')$.

3. Cliques and efficient graph.

3.1. Nondecreasing $\mathcal{Q}$.

In this section, we focus on maps $\mathcal{Q}$ which are compatible with the partial order $\preccurlyeq$ on the graphs in $\mathcal{G}_N$. This will allow us to give conditions for comparing the value of two graphs independently of the system (2.1) and for characterizing the efficient graph rigorously.

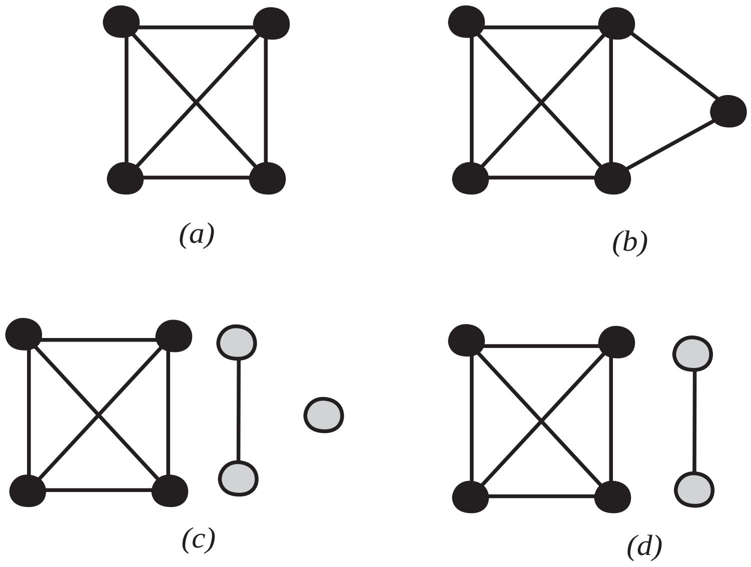


Fig. 3.1. *Some examples of graphs. All but* (b) *are clique graphs.*

We start with the following simple result.

Proposition 3.1.

(i) *If $g \preccurlyeq g'$, then for all $Q > 0$, $J^\star_{g'}(Q) \le J^\star_g(Q)$.*

(ii) *If $Q \le Q'$, then for all $g \in \mathcal{G}_N, J^\star_g(Q) \le J^\star_g(Q')$.*

We thus obtain the following immediate corollary.

Corollary 3.2. *If the map $\mathcal{Q}$ is nondecreasing and $g \preccurlyeq g'$, $J^\star_{g'}(\mathcal{Q}(g)) \le V(g')$.*

Proof. Applying item (ii) of Proposition 3.1 to $Q = \mathcal{Q}(g)$ and $Q' = \mathcal{Q}(g')$ yields

$$J^\star_{g'}(\mathcal{Q}(g)) \le J^\star_{g'}(\mathcal{Q}(g')).$$

Since by definition

$$J^\star_{g'}(\mathcal{Q}(g')) = V(g'),$$

we immediately obtain

$$J^\star_{g'}(\mathcal{Q}(g)) \le V(g'). \qquad \square \tag{3.1}$$

Before we go further and prove our first result on graph efficiency, we need to introduce the following concept borrowed from [12].

Definition 3.3. *A graph $g \in \mathcal{G}_N$ is said to be a* clique graph *if each of its connected components is a clique, i.e., a complete subgraph.*

Examples and counterexamples of clique graphs are given in Figure 3.1. As we will see, these graphs are useful because their value can be readily computed.

Proposition 3.4. *Let $g \in \mathcal{G}_N$ be a clique graph and the map $\mathcal{Q}$ be structure compatible. Then $V(g) = J^\star(\mathcal{Q}(g))$. In particular, $V(g) = \lambda_{\max}(P(g))$, where $P(g)$*

is the unique positive definite solution of the Riccati equation

$$P(g) = A^*P(g)A + \mathcal{Q}(g) - A^*P(g)B\left(B^*P(g)B + I\right)^{-1}B^*P(g)A. \tag{3.2}$$

Proof. We have already mentioned in the previous section that the unstructured control law minimizing $J(\mathcal{Q}(g), .)$ and the corresponding optimal value are given by the solution to Riccati equation (3.2). Since g is a clique graph, using a reordering of the vertices, the matrix $\mathcal{A}(g)$ can be converted to a block-diagonal form. The systems thus become decoupled and one can readily show that, under our assumptions, the optimal control law $K(g)$, in fact, has structure g (see, e.g., the proof of Proposition 2.6 in [10]).

Now, note that, by definition of $K(g)$,

$$J(\mathcal{Q}(g), x_0; u) \geq J(\mathcal{Q}(g), x_0; K(g))$$

for all x_0 and arbitrary control laws u. Hence, for all u,

$$J(\mathcal{Q}(g), u) \geq J(\mathcal{Q}(g), K(g))$$

and, in particular, considering control laws with structure g,

$$V(g) = J^\star_g(\mathcal{Q}(g)) \geq J(\mathcal{Q}(g), K(g)).$$

The fact that $K(g)$ has structure g then completes the proof. □

Proposition 3.4 and its proof are reminiscent of the results of [2], where it is shown that, for spatially invariant systems, the optimal controller is itself spatially invariant. Here, the optimal controller has the same structure as the clique graph, when the map $\mathcal{Q}$ is structure compatible. This property allows us to compare the value of a clique graph to that of its supergraphs.

THEOREM 3.5. *Let $\mathcal{Q}$ be nondecreasing and structure compatible and g be a clique graph. Then $V(g) \leq V(g')$ for all $g' \in \mathcal{G}_N$ such that $g \preccurlyeq g'$. In particular, the graph $g^\star$ characterized by $E(g^\star) = \{(i,i) : 1 \leq i \leq N\}$ is efficient, as defined in* (2.9). *In other words, the minimal control topology* (*for cost* (2.3)) *is fully decentralized.*

Proof. By Proposition 3.4, since g is a clique graph and $\mathcal{Q}$ is structure compatible, $V(g) = J^\star(\mathcal{Q}(g))$. Also, since $\mathcal{Q}$ is nondecreasing, we can use relation (3.1) to write $J^\star_{g'}(\mathcal{Q}(g)) \leq V(g')$. Finally, by definition of the various minimization problems, we have

$$V(g) = J^\star(\mathcal{Q}(g)) \leq J^\star_{g'}(\mathcal{Q}(g)) \leq V(g').$$

That the fully decentralized topology is minimal follows from the fact that $g^\star$, as defined in the theorem, is a clique graph and, clearly, that $g^\star \preccurlyeq g$ for all $g \in \mathcal{G}_N$. □

The result of Theorem 3.5 illustrates that, if a cost is charged for communication, cooperation can sometimes be detrimental. Note that from [11] we know that for the purposes of stabilizability and controllability, all communication topologies are equivalent. However, different information patterns yield different performance and this result states that more communication links (and hence less constraints on the structure of the control law) may not automatically translate into better performance in the case of structure compatibility.

3.2. Merely structure-compatible $\mathcal{Q}$. The previous section has shown that it is possible to compare clique graphs to their supergraphs when map $\mathcal{Q}$ is structure compatible and nondecreasing, and that adding edges to a clique graph is then always detrimental to the value. These results hold independently of the subsystems (2.1) under consideration.

In this section, we remove the nondecreasing assumption on map $\mathcal{Q}$ and derive conditions for comparing clique graphs to *arbitrary* graphs. In the remainder of this section, we will assume that either

- g is a clique graph and map $\mathcal{Q}$ is structure compatible, or
- g is the complete graph, i.e., $E(g) = \{(i,j),\ 1 \le i \le j \le N\}$, and map $\mathcal{Q}$ is arbitrary.

THEOREM 3.6. *Let $\gamma = \lambda_{\max}(P(g))$, where $P(g)$ is the solution of Riccati equation (2.4) with $Q = \mathcal{Q}(g)$. A graph $g' \in \mathcal{G}_N$ satisfies $V(g') \le \gamma = V(g)$ if and only if $\lambda^\star \le \gamma$, where $\lambda^\star$ is defined as*

$$\lambda^\star := \begin{cases} & \min\limits_{P',K'} \lambda_{\max}(P') \\ \textit{subject to} & P' \ge (A+BK')^* P'(A+BK') + \mathcal{Q}(g') + K'^* K', \\ & K' \in \mathcal{K}_{m,n}(g'). \end{cases} \tag{3.3}$$

Proof. Assume $\lambda^\star \le \gamma$. Then there exists $P' > 0$ and $K' \in \mathcal{K}_{m,n}(g')$ such that (3.3) holds and $\lambda_{\max}(P') \le \gamma$. Also, we see that along any closed-loop trajectory with initial condition x_0,

$$x^*(k)P'x(k) - x^*(k+1)P'x(k+1) \ge x^*(k)\mathcal{Q}(g')x(k) + u^*(k)u(k).$$

Then, summing over k, we obtain

$$x_0^* P' x_0 \ge \sum_{k=0}^{\infty} \left(x^*(k)\mathcal{Q}(g')x(k) + u^*(k)u(k)\right) = J_{g'}\left(\mathcal{Q}(g'), x_0; u\right) \tag{3.4}$$

for the control law u satisfying $u(k) = K'x(k)$. Taking the supremum over x_0 over the unit ball yields

$$\gamma \ge J(\mathcal{Q}(g'), K') \ge V(g').$$

Reciprocally, assume $V(g') \le V(g)$. In this case, $V(g')$ is bounded, and thus there exists a stabilizing control law with structure g'. By virtue of Proposition 2.2, this implies that there exists $\bar{K}'$ such that $V(g') = J(\mathcal{Q}(g'), \bar{K}')$. The control law corresponding to $\bar{K}'$ must itself be stabilizing, and thus, according to Proposition 2.1,

$$V(g') = \lambda_{\max}(\bar{P}'),$$

where $\bar{P}'$ is the positive definite solution of Lyapunov equation (2.6) with $Q = \mathcal{Q}(g')$.

Since $(\bar{P}', \bar{K}')$ is feasible for the optimization problem on the right-hand side of (3.3), this problem has the same optimal value as the problem $(\mathcal{P})$,

$$(\mathcal{P}) \quad \begin{cases} & \inf\limits_{P',K'} \lambda_{\max}(P') \\ \text{subject to} & P' \ge (A+BK')^* P'(A+BK) + \mathcal{Q}(g') + K'^* K', \\ & K' \in \mathcal{K}_{m,n}(g'), \\ & \lambda_{\max}(P') \le \lambda_{\max}(\bar{P}'). \end{cases}$$

The feasible set of $(\mathcal{P})$ is compact since it is clearly closed and, if (P, K) is feasible for $(\mathcal{P})$, both $\lambda_{\max}(P)$ and $\sigma_{\max}(K)$ are bounded by $\lambda_{\max}(\bar{P}')$. Hence the infimum in $(\mathcal{P})$ is attained; i.e., there exists (P_0, K_0) such that the optimal value of problem $(\mathcal{P})$ is $\lambda_{\max}(P_0)$. Then, clearly, $\lambda^\star = \lambda_{\max}(P_0)$, which means that the optimal value is also attained in the problem on the right-hand side of (3.3) and, by definition of P_0,

$$\lambda^\star = \lambda_{\max}(P_0) \leq \lambda_{\max}(\bar{P}') = V(g') \leq \gamma. \qquad \square$$

Condition (3.3), although both necessary and sufficient, is not practical for comparing the value of a graph to that of a clique graph, since problem (3.3) is not easily solved numerically. This is because it involves a bilinear constraint on variables P' and K', and the change of variables traditionally used to convexify static state feedback synthesis problems is inoperative here because of the structural constraint on K'. Some approaches have been proposed recently for obtaining merely sufficient, but *convex*, synthesis conditions in a similar context, from restricting oneself to a diagonal matrix P' [1] to introducing additional variables [7]. These methods could be applied to the present problem as well, and used to derive a computable upper bound for $\lambda^\star$ and, in turn, a sufficient condition for g' having lower value than g. However, we give different sets of convex sufficient and necessary conditions for comparing the values of graphs.

THEOREM 3.7 (sufficient condition). *Let $P(g)$ be the unique positive definite solution of Riccati equation* (3.2) *with $Q = \mathcal{Q}(g)$. If there exists a matrix $K \in \mathcal{K}_{m,n}(g')$ such that the following linear matrix inequality (LMI) in matrix variable K is satisfied:*

$$\begin{bmatrix} -P(g) + \mathcal{Q}(g') & \begin{bmatrix} (A+BK)^* & K^* \end{bmatrix} \\ \begin{bmatrix} (A+BK) \\ K \end{bmatrix} & -\begin{bmatrix} P(g)^{-1} & 0 \\ 0 & I \end{bmatrix} \end{bmatrix} < 0, \tag{3.5}$$

then $V(g) \geq V(g')$.

Proof. If matrix K is feasible for LMI (3.5), the pair $(P(g), K)$ is feasible for problem (3.3) since, according to the Schur complement formula,

$$-P(g) + \mathcal{Q}(g') + (A+BK)^* P(g)(A+BK) + K^*K < 0.$$

As a result, $\lambda^\star \leq \lambda_{\max}(P(g)) = \gamma$ and, according to Theorem 3.6, $V(g') \leq V(g)$. $\square$

THEOREM 3.8 (necessary condition). *If $V(g) \geq V(g')$, then there exists a matrix $K \in \mathcal{K}_{m,n}(g')$ and a matrix $X > 0$ such that the following LMI (in K and X) is satisfied:*

$$\begin{bmatrix} -\gamma I + \mathcal{Q}(g') & \begin{bmatrix} (A+BK)^* & K^* \end{bmatrix} \\ \begin{bmatrix} (A+BK) \\ K \end{bmatrix} & -\begin{bmatrix} X & 0 \\ 0 & I \end{bmatrix} \end{bmatrix} \leq 0, \tag{3.6}$$

where γ is defined as in Theorem 3.6.

Proof. By Theorem 3.6, if $V(g) \geq V(g')$, there exists (P', K') feasible for problem (3.3) with $\lambda_{\max}(P') \leq \gamma$. Hence, there exist P' and K' such that $P' \leq \gamma I$ and $P' \geq (A+BK')^* P'(A+BK') + \mathcal{Q}(g') + K'^*K'$ and, as a result,

$$\gamma I \geq (A+BK')^* P'(A+BK') + \mathcal{Q}(g') + K'^*K'.$$

Using a Schur complement and letting $X = P'$ shows that the LMI of Theorem 3.8 is feasible. $\square$

4. Comparing arbitrary graphs. In this section, we extend some of the tools presented so far and compare the values of arbitrary graphs for any given map $\mathcal{Q}$. Because it is already difficult to compute the value of an arbitrary graph (as opposed to the case of clique graphs and structure-compatible maps), we settle for tractable sufficient conditions on the weighting matrices $\mathcal{Q}(g)$ and $\mathcal{Q}(g')$, which allow us to compare graphs g and g'. More precisely, we ask what (tractably testable) properties of a map $\mathcal{Q}$ are sufficient to guarantee that one graph has a smaller value than another one.

THEOREM 4.1 (sufficient condition). *Consider two graphs g and g'. If there exist $K \in \mathcal{K}_{m,n}(g')$ and a matrix $P > 0$ such that*

$$P = (A+BK)^*P(A+BK) + \mathcal{Q}(g') + K^*K, \tag{4.1}$$

$$\mathcal{Q}(g) \geq \mathcal{Q}(g') + \left(K + S^{-1}B^*PA\right)^* S \left(K + S^{-1}B^*PA\right), \tag{4.2}$$

$$S = B^*PB + I, \tag{4.3}$$

then $V(g) > V(g')$.

Proof. Assume $P > 0$ satisfies (4.1). Then, since $\mathcal{Q}(g') + K^*K > 0$, by the properties of a discrete algebraic Lyapunov equation [9], the matrix $(A+BK)$ is Schur. Thus,

$$\lim_{k\to\infty} x(k) = 0$$

for the closed-loop system

$$x(k+1) = (A+BK)x(k),$$

starting from any initial condition x_0. Also, proceeding as in the proof of Theorem 3.6, we see that along any closed-loop trajectory with initial condition x_0,

$$x_0^*Px_0 = J_{g'}\left(\mathcal{Q}(g'), x_0; u\right) \tag{4.4}$$

for the control law u satisfying $u(k) = Kx(k)$. Now from (4.1) we see that P satisfies

$$P = A^*PA - A^*PBS^{-1}B^*PA + \mathcal{Q}(g') + \left(K + S^{-1}B^*PA\right)^* S \left(K + S^{-1}B^*PA\right),$$

where $S = B^*PB + I$. This Riccati equation is identical to the one obtained using LQ control theory if we were to look for an *unstructured* control law that minimizes a cost function of the form (2.3), but with the weighting matrix

$$Q = \mathcal{Q}(g') + \left(K + S^{-1}B^*PA\right)^* S \left(K + S^{-1}B^*PA\right).$$

Thus, for all initial conditions x_0 and control laws v,

$$x_0^*Px_0 \leq J(\mathcal{Q}(g') + \left(K + S^{-1}B^*PA\right)^* S \left(K + S^{-1}B^*PA\right), x_0; v)$$

and, maximizing over x_0,

$$J_{g'}(\mathcal{Q}(g'), u) \leq J(\mathcal{Q}(g') + \left(K + S^{-1}B^*PA\right)^* S \left(K + S^{-1}B^*PA\right), v)$$

for all control laws v. But, from (4.2), we obtain that

$$J(\mathcal{Q}(g') + \left(K + S^{-1}B^*PA\right)^* S \left(K + S^{-1}B^*PA\right)) \leq J(\mathcal{Q}(g), v)$$

and, thus, that $J_{g'}(\mathcal{Q}(g'), u) \leq J(\mathcal{Q}(g), v)$ for all v. In particular, this implies that

$$V(g') \leq J_{g'}(\mathcal{Q}(g'), u) \leq V(g). \qquad \square$$

Conditions (4.1) can be used to *design* a map $\mathcal{Q}$ to enforce some desired control topology. For example, imagine a situation where the map $\mathcal{Q}$ is chosen by one institution (the price designer) while the controller is synthesized by another (the network builder) and where a topology, g', has been agreed on and implemented. Then, the price designer can ensure that there is no incentive to build a new edge by choosing a stabilizing control gain $K \in \mathcal{K}_{m,n}(g')$, solving Lyapunov equation (4.1) and picking $\mathcal{Q}$ such that (4.2) holds. On the other hand, the network builder is only given the map $\mathcal{Q}$. If it wants to use conditions (4.1) to determine whether it is advantagous to add new edges to a pre-existing control topology (or, more aptly, to find a certificate that it is detrimental to do so) it has to solve equations (4.1) for both P and K. Even after using the Schur complement formula on inequality (4.2) and rewriting (4.1) as two matrix inequalities, this is still a hard task to perform, since one then has to solve a set of bilinear matrix inequalities which, in general, is NP-hard [22].

Another type of sufficient condition can be obtained for comparing arbitrary graphs by building on the ideas of section 3. In particular, we can state the following.

THEOREM 4.2. *Let g, g' be two graphs and $\mathcal{Q}$ be any weight map. If there exists a diagonal matrix Λ such that $\mathcal{Q}(g) < \Lambda < \mathcal{Q}(g')$, then $V(g) < V(g')$.*

Proof. Let e be the fully decentralized graph, i.e., $E(e) = \{(i,i) : 1 \leq i \leq N\}$, and define $\mathcal{H}(g) := \Lambda$ for all g. Map $\mathcal{H}$ is clearly structure compatible (and nondecreasing), and we can compute the value of any graph h in $\mathcal{G}_N$ by using this weighting map instead of $\mathcal{Q}$. We will denote this value by $V_{\mathcal{H}}(h)$. We can now proceed in two steps:

(a) Using Proposition 3.1, we see that

$$\begin{aligned} V(g) &= J_g^\star(\mathcal{Q}(g)) \\ &\leq J_e^\star(\mathcal{Q}(g)) \text{ since } g \succcurlyeq e \\ &\leq J_e^\star(\Lambda) = V_{\mathcal{H}}(e). \end{aligned}$$

(b) Using Theorem 3.5, we see that $V_{\mathcal{H}}(e) \leq V_{\mathcal{H}}(g')$ since $\mathcal{H}$ is structure compatible. But $\mathcal{H}(g') = \Lambda \leq \mathcal{Q}(g')$, and so

$$J_{g'}^\star(\mathcal{H}(g')) \leq J_{g'}^\star(\mathcal{Q}(g')).$$

Hence $V(g) \leq V_{\mathcal{H}}(e) \leq V_{\mathcal{H}}(g') \leq V(g')$. $\square$

In fact, the arguments of the previous proof are still valid if matrix Λ, instead of being diagonal, has the structure of a clique graph. We can thus state the following.

COROLLARY 4.3. *Let h be a clique subgraph of both graphs g and g'. If there exists a matrix $M \in \mathcal{K}_{m,n}(h)$ such that $\mathcal{Q}(g) < M < \mathcal{Q}(g')$, then $V(g) < V(g')$.*

5. Communication topology design for multivehicle systems: An example.

In this section, we illustrate our results by applying them to the problem of determining the most economical information pattern required to keep multiple vehicles in a geometric formation, while taking into account the cost of communication.

Consider the system pictured in Figure 5.1 and composed of three independent nonholonomic vehicles modeled, in continuous time, by the equations

$$\left\{ \begin{aligned} \dot{\mathrm{x}}_i &= V_i \cos\theta_i, \\ \dot{\mathrm{y}}_i &= V_i \sin\theta_i, \\ \dot{\theta}_i &= \omega_i, \end{aligned} \right. \tag{5.1}$$

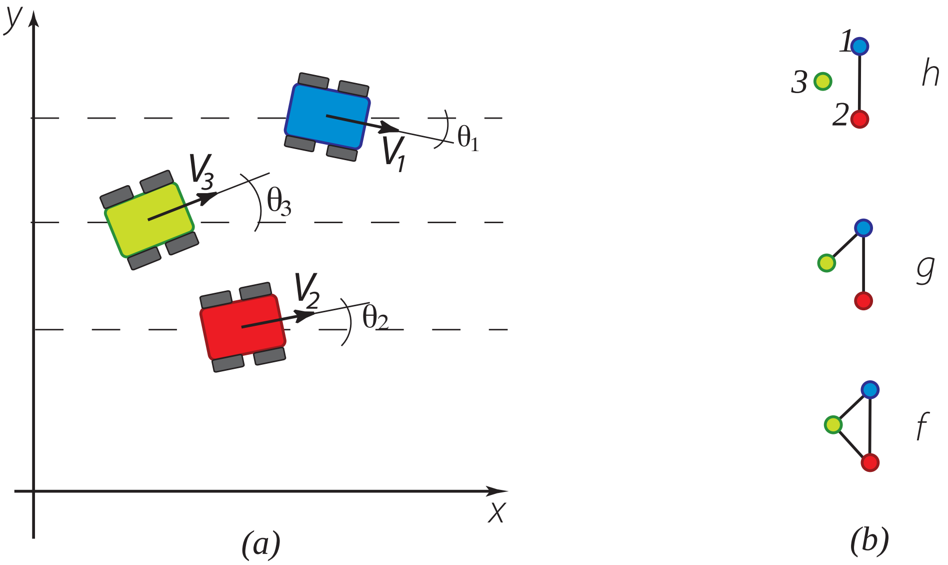


FIG. 5.1. *Multivehicle formation.* (a) *Kinematics of the problem.* (b) *Three possible communication topologies. See text for details.*

for all $i = 1, \dots, 3$, where $(\mathrm{x}_i, \mathrm{y}_i)$ are the coordinates of the center of mass of the ith vehicle, θ_i is its heading angle, V_i is its velocity, and ω_i (the control input) is its angular velocity. Linearizing these equations around straight-line trajectories parallel to the x-axis (i.e., around $\theta_i \equiv 0$ and $y_i \equiv$ constant, for each i) and discretizing them with a zeroth-order hold scheme with a time step of 0.1 yields the double-integrator model

$$\begin{pmatrix} \delta y_i(k+1) \\ \theta_i(k+1) \end{pmatrix} = \begin{pmatrix} 1 & 0.1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \delta y_i(k) \\ \theta_i(k) \end{pmatrix} + \begin{pmatrix} 0.005 \\ 0.1 \end{pmatrix} \omega_i(k) \tag{5.2}$$

for each vehicle, where we have set $V_i \equiv 1$ for all i, and δy_i designates the deviation from the desired constant value of y_i. Equations (5.2) can also be thought of as capturing the motion of the vehicles in a frame moving parallel to the x-axis with unit velocity. We want these vehicles to evolve in formation so that

- angle $\theta_i(k)$ and deviation $\delta y_i(k)$ remain small for all i, at each time k;
- vehicle number 3 is always at the middle point between vehicle 1 and 2.

With such requirements, it makes sense to design control laws $\{\omega_i\}_{i=1,\dots,3}$ for subsystems (5.2) so that the cost function

$$\sum_{k=0}^{\infty} \left(\delta y_1^2(k) + \delta y_2^2(k) + \delta y_3^2(k)\right) + \left(\theta_1^2(k) + \theta_2^2(k) + \theta_3^2(k)\right) \\ + \left(\delta y_3(k) - \frac{\delta y_1(k) + \delta y_2(k)}{2}\right)^2 + \left(\omega_1^2(k) + \omega_2^2(k) + \omega_3^2(k)\right)$$

is minimized in closed-loop form. This amounts to solving an optimal control problem of the form (2.5) with matrix Q_0 given by

$$Q_0 = \left(\begin{array}{cc|cc|cc} \frac{5}{4} & 0 & \frac{1}{2} & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ \hline \frac{1}{2} & 0 & \frac{5}{4} & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ \hline -1 & 0 & -1 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right). \tag{5.3}$$

In (5.3), matrix Q_0 is partitioned conformably to the subsystems, with the first and second coordinates of subsystem i's state being δy_i and θ_i, respectively.

Note that without any communication cost, the optimal controller for the matrix Q_0 is full, which means that the optimal graph is a full graph. In order to determine an efficient communication graph for this problem, we must define a communication cost in the form of a mapping $\mathcal{Q} : \mathcal{G}_3 \to \mathbb{S}^6$. We choose to adopt an edge-separable map with externalities, such that

$$\mathcal{Q}(g) = Q_0 + \sum_{(i,j)\in E(g)} P_{ij} \text{ for all } g,$$

with each positive definite matrix P_{ij} having all its blocks equal to zero except for the diagonal and the (i, j) and (j, i) ones. Note that the map $\mathcal{Q}$ so obtained is nondecreasing but *not* structure compatible, since Q_0 is not block diagonal. We want to determine price matrices such that the graphs pictured in Figure 5.1(b) satisfy $V(g) \le V(f)$ and $V(h) \le V(g)$. We start by choosing

$$P_{12} = \left(\begin{array}{cc|cc|cc} 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right)$$

and use Theorem 4.1 to find a price matrix P_{13} such that $V(h) \le V(g)$. Picking the stabilizing controller

$$K = \left(\begin{array}{cc|cc|cc} -1 & -1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & -1 & -1 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & -1 & -1 \end{array}\right)$$

with structure h, one can show that (4.1) are satisfied for some matrix $P > 0$ if we take

$$P_{13} = \mathcal{Q}(g) - \mathcal{Q}(h) = \left(\begin{array}{cc|cc|cc} 10 & 0 & 0 & 0 & -0.5 & 0 \\ 0 & 5 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 10 & 1 & 0 & 0 \\ 0 & 0 & 1 & 10 & 0 & 0 \\ \hline -0.5 & 0 & 0 & 0 & 10 & 0 \\ 0 & 0 & 0 & 0 & 0 & 5 \end{array}\right).$$

To find a matrix P_{23} such that $V(f) \ge V(g)$, we can use Theorem 4.2 and solve the two LMIs

$$\begin{aligned} Q_0 + P_{12} + P_{13} &< \Lambda, \\ \Lambda &< Q_0 + P_{12} + P_{13} + P_{23} \end{aligned}$$

in the structured variables Λ and P_{23}. Using SeDuMi, we find that the following pair is feasible:

$$P_{23} = \left(\begin{array}{cc|cc|cc} 78.797 & 0 & 0 & 0 & 0 & 0 \\ 0 & 78.797 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 78.797 & 0 & 13.1328 & 0 \\ 0 & 0 & 0 & 78.797 & 0 & 13.1328 \\ \hline 0 & 0 & 13.1328 & 0 & 78.797 & 0 \\ 0 & 0 & 0 & 13.1328 & 0 & 78.797 \end{array}\right); \ \Lambda = 39.3985\, I_6.$$

By adopting the price matrices given above, the price designer can thus ensure that graph h will be chosen by the network builder.

6. Conclusions and future work.

In this paper, we proposed and partially analyzed a new model for determining the influence of the structure of a controller on closed-loop performance in distributed control design problems. For a plant composed of dynamically uncoupled subsystems, we proposed making one of the weight matrices of the classical LQR cost function topology dependent, so as to capture the cost of communication between subsystems. For some models of such dependencies, we investigated the existence and properties of an optimal structured controller.

This paper is only a first attempt at studying the influence of interconnection topology on performance in distributed control design problems. The problem is hard because the underlying problem of determining optimal structured controllers is hard. The chief virtue of our approach is that it allows us to circumvent this problem and enables us to give rigorous statements about optimal topologies and comparing topologies to each other. Our results thus complement the more heuristic claims of [3] and [21].

Even within the realm of this restricted model, many unanswered questions remain. For example, it would be nice to be able to compare the values of any two graphs and not just of those that are comparable for the partial order $\preccurlyeq$. Likewise, one may ask whether it is possible to derive bounds similar to those of Theorem 4.1 that are tractable or explicit. Another avenue of future research would be to make the communication cost model more physically motivated by considering metrics such as entropy rate. Finally, the effect of imperfections in the communication links also needs to be studied.

Acknowledgments. Discussions with Profs. Anders Rantzer and Tudor Stoenescu, on this and related topics, are gratefully acknowledged.

REFERENCES

[1] G. Ayres de Castro and F. Paganini, *Convex synthesis of localized controllers for spatially invariant systems*, Automatica, 38 (2002), pp. 445–456.

[2] B. Bamieh, F. Paganini, and M. A. Dahleh, *Distributed control of spatially invariant systems*, IEEE Trans. Automat. Control, 47 (2002), pp. 1091–1107.

[3] G. M. Belanger, S. Ananyev, J. L. Speyer, D. F. Chichka, and J. R. Carpenter, *Decentralized control of satellite clusters under limited communication*, AIAA J. Guidance Control Dynamics, 29 (2006), pp. 134–145.

[4] R. Bellman, *Large systems*, IEEE Trans. Automat. Control, 19 (1974), p. 465.

[5] V. Blondel and J. Tsitsiklis, *NP-hardness of some linear control design problems*, SIAM J. Control, 35 (1997), pp. 2118–2127.

[6] J. Crémer, *A partial theory of the optimal organization of a bureaucracy*, Bell J. Econom., 11 (1981), pp. 683–693.

[7] M. C. de Oliveira, J. C. Geromel, and J. Bernussou, *Extended H_2 and H_∞ norm characterizations and controller parametrizations for discrete-time systems*, Internat. J. Control, 75 (2002), pp. 666–679.

[8] W. B. Dunbar and R. M. Murray, *Distributed receding horizon control with applications to multi-vehicle formation stabilization*, Automatica J. IFAC, 42 (2006), pp. 549–558.

[9] Z. Gajic and M. T. J. Qureshi, *Lyapunov Matrix Equation in System Stability and Control*, Academic Press, New York, 1995.

[10] V. Gupta, *Distributed Estimation and Control in Networked Systems*, Ph.D. thesis, Department of Electrical Engineering, California Institute of Technology, Pasadena, CA, 2007.

[11] V. Gupta, B. Hassibi, and R. M. Murray, *A sub-optimal algorithm to synthesize control laws for a network of dynamic agents*, Internat. J. Control, 78 (2005), pp. 1302–1313.

[12] F. Harary, *Graph Theory*, Addison-Wesley, New York, 1994.

[13] Y. C. Ho and K. Chu, *Team decision theory and information structures in optimal control problems–Part I*, IEEE Trans. Automat. Control, 17 (1972), pp. 15–22.

[14] M. O. Jackson, *A survey of network formation models: Stability and efficiency*, in Group Formation in Economics: Networks, Clubs, and Coalitions, G. Demange and M. Wooders, eds., Cambridge University Press, Cambridge, UK, 2004, pp. 11–57.

[15] C. Langbort, R. S. Chandra, and R. D'Andrea, *Distributed control design for systems interconnected over an arbitrary graph*, IEEE Trans. Automat. Control, 49 (2004), pp. 1502–1519.

[16] X. Liu and A. J. Goldsmith, *Wireless network design for distributed control*, in Proceedings of 43rd IEEE Conference on Decision and Control, Paradise Island, Bahamas, 2004, pp. 2823–2829.

[17] J. Marschak and R. Radner, *Economic Theory of Teams*, Cowles Foundation Monograph 22, Yale University Press, New Haven, CT, 1972.

[18] N. Motee and A. Jadbabaie, *Optimal control of spatially distributed systems*, in Proceedings of the IEEE American Control Conference, New York, NY, 2007, pp. 778–783.

[19] M. Rotkowitz and S. Lall, *A characterization of convex problems in decentralized control*, IEEE Trans. Automat. Control, 51 (2006), pp. 274–286.

[20] N. Sandell and M. Athans, *Solution of some nonclassical LQG stochastic decision problems*, IEEE Trans. Automat. Control, 19 (1974), pp. 108–116.

[21] R. Smith and F. Hadaegh, *Closed-loop dynamics of cooperative vehicle formations with parallel estimators and communication*, IEEE Trans. Automat. Control, 52 (2007), pp. 1404–1414.

[22] O. Toker and H. Özbay, *On the NP-hardness of solving bilinear matrix inequalities and simultaneous stabilization with static output feedback*, in Proceedings of the IEEE American Control Conference, Seattle, WA, 1995, pp. 2525–2526.

[23] T. Van Zandt, *Decentralized information processing in the theory of organizations*, in Contemporary Economic Issues, Vol. 4: Economic Design and Behavior, M. Sertel, ed., MacMillan Press, London, 1999, pp. 125–160.

[24] S. Wang and E. J. Davison, *On the stabilization of decentralized control systems*, IEEE Trans. Automat. Control, 18 (1973), pp. 473–478.

[25] H. S. Witsenhausen, *A counterexample in stochastic optimum control*, SIAM J. Control, 6 (1968), pp. 131–147.

[26] L. Xiao and S. Boyd, *Fast linear iterations for distributed averaging*, Systems Control Lett., 53 (2004), pp. 65–78.

SIAM J. Control Optim.
Vol. 48, No. 2, pp. 415–437

© 2009 Society for Industrial and Applied Mathematics

EXISTENCE RESULTS FOR OPTIMAL CONTROL PROBLEMS WITH SOME SPECIAL NONLINEAR DEPENDENCE ON STATE AND CONTROL*

PABLO PEDREGAL† and JORGE TIAGO†

Abstract. We present a general approach to prove existence of solutions for optimal control problems not based on typical convexity conditions, which quite often are very hard, if not impossible, to check. By taking advantage of several relaxations of the problem, we isolate an assumption which guarantees the existence of solutions of the original optimal control problem. To show the validity of this crucial hypothesis through various means and in various contexts is the main goal of this paper. In each such situation, we end up with some existence result. In particular, we would like to stress a general result that takes advantage of the particular structure of both the cost functional and the state equation. One main motivation for our work here comes from a model for guidance and control of ocean vehicles. Some explicit existence results and comparison examples are given.

Key words. relaxation, moments, nonlinear, nonconvex

AMS subject classifications. 49J15, 49M20, 78M05

DOI. 10.1137/08071805X

1. Introduction.

This paper focuses on the analysis of optimal control problems of the general form

$$(P_1)\quad \begin{cases} \text{Minimize in } u: & \displaystyle\int_0^T \left[\sum_{i=1}^{s} c_i(x(t))\phi_i(u(t))\right] dt \\ \text{subject to} & x'(t) = \displaystyle\sum_{i=1}^{s} Q_i(x(t))\phi_i(u(t)) \text{ in } (0,T), \quad x(0) = x_0 \in R^N, \\ \text{and} & u \in L^\infty(0,T), \quad u(t) \in K, \end{cases}$$

where $K \subset R^m$ is compact. The state $x : (0,T) \to R^N$ takes values in R^N.

The mappings

$$c_i : R^N \to R, \quad \phi_i : R^m \to R, \quad Q_i : R^N \to R^N$$

as well as the restriction set $K \subset R^m$ will play a fundamental role. We assume, at this initial stage, that c_i are continuous, ϕ_i are of class $\mathcal{C}^1$, and each Q_i is Lipschitz so that the state system is well-posed.

In such a general form, we cannot apply results for nonnecessarily convex problems like the ones in [2], [6], [25], or [30]. Besides, techniques based on Bauer's maximum principle (see [3], [4]) are quite difficult to extend to our general setting because it is hard to analyze the concavity of the cost functional when the dependence on both

*Received by the editors March 10, 2008; accepted for publication (in revised form) August 18, 2008; published electronically February 13, 2009.
http://www.siam.org/journals/sicon/48-2/71805.html

†Universidad de Castilla-La Mancha, ETSI Industriales, 13071 Ciudad Real, Spain (pablo.pedregal@uclm.es, jorge.tiago@uclm.es). The first author's work was supported by the research projects MTM2007-62945 of the MCyT (Spain) and PCI08-0084-0424 of the JCCM (Castilla-La Mancha). The second author's research was supported by Fundação para a Ciência e a Tecnologia (Portugal), by the doctoral degree grant SFRH-BD-22107-2005, and by CMAF-Lisbon University through FEDER and FCT-Plurianual 2007.

state and control comes in product form. Also Rockafellar's variational reformulation introduced in [27], and well described in [7], [12] or recently in [23] or [24], looks as if it cannot avoid assuming a separated dependence on the state and control variables, since this is the structure of the variational problem for which the existence of solution has been so far ensured [8].

Concerning the classical Filippov–Roxin theory introduced in [13] and [31], it is not easy at all to know if typical convexity assumptions hold, or when they may hold, as we can see from the examples and counterexamples in [7]. When analyzing explicit examples, one realizes such difficulties coming from the need of a deep understanding of typical orientor fields. The same troubles would arise when applying refinements of this result as the ones in [20] and [21].

Recently (see [9]), an existence result has been shown for minimum-time problems where the typical convexity assumptions over the set-valued function on the differential inclusion have been replaced by more general conditions. In fact, the intersection of this result with the ones we present here is not empty, although, as we will comment, our frame extends to situations not covered by this result. Such analysis can be done by writing problem (P_1) as a minimum-time problem as suggested in [7].

Our aim is to provide hypotheses on the different ingredients of the problem so that existence of solutions can be achieved through an independent road. Actually, it is not easy to claim whether our results improve on classical or more recent general results. They provide an alternative tool which can be more easily used in practice than such results when one faces an optimal control problem under the special structure we consider here. As a matter of fact, convexity will also occur in our statements but in an unexpected and nonstandard way.

Before stating our main general result, a bit of notation is convenient. We will write

$$c : R^N \to R^s, \quad \phi : R^n \to R^s, \quad Q : R^N \to R^{Ns}, \tag{1}$$

with components c_i, ϕ_i, and Q_i, respectively. Consider also a new ingredient of the problem related to ϕ. Suppose that there is a $\mathcal{C}^1$ mapping

$$\Psi : R^s \to R^{s-n}, \quad \Psi = (\psi_1, \dots, \psi_{s-n}) \quad (s > n), \tag{2}$$

so that $\phi(K) \subset \{\Psi = 0\}$. This is simply saying, in a rough way, that the embedded (parametrized) manifold $\phi(K)$ of R^s is part of the manifold defined implicitly by $\Psi = 0$. In practical terms, it suffices to check that the composition $\Psi(\phi(u)) = 0$ for $u \in K$.

For a pair (c, Q), put

$$\mathcal{N}(c, Q) = \{v \in R^s : Qv = 0, cv \leq 0\}. \tag{3}$$

Similarly, set

$$\begin{aligned} \mathcal{N}(K, \phi) = \big\{ & v \in R^s : \text{ for each } u \in K, \\ & \text{either } \nabla\Psi(\phi(u))v = 0 \text{ or } \exists i \text{ s.t. } \nabla\psi_i(\phi(u))v > 0 \big\}. \end{aligned} \tag{4}$$

Our main general result is the following.

THEOREM 1.1. *Assume that the mapping Ψ as above is strictly convex (componentwise) and $\mathcal{C}^1$. If for each $x \in R^N$ we have*

$$\mathcal{N}(c(x), Q(x)) \subset \mathcal{N}(K, \phi), \tag{5}$$

then the corresponding optimal control problem (P_1) *admits at least one solution.*

As it stands, this result looks rather abstract, and it is hard to grasp to what extent it may be applied in more specific situations.

A particular, yet still under some generality, situation where this result can be implemented is the case of polynomial dependence where the ϕ_i's are polynomials of various degrees. The main structural assumption, in addition to the one coming from the set K, is concerned with the convexity of the corresponding mapping Ψ.

Suppose we take $\phi_i(u) = u_i$, $i = 1, 2, \dots, n$, and $\phi_{n+i}(u)$, $i = 1, 2, \dots, s-n$, convex polynomials of whatever degree, or simply polynomials whose restriction to K is convex. In particular, K itself is supposed to be convex. Then we can take

$$\Psi_i(v) = \phi_{n+i}(\overline{v}) - v_{n+i}, \quad i = 1, 2, \dots, s-n, \quad \overline{v} = (v_i)_{i=1,2,\dots,n}. \tag{6}$$

In this case, it is clear that

$$\Psi(\phi(u)) = 0 \text{ for } u \in K,$$

by construction, and, in addition, Ψ is smooth and convex. The important constraint (5) can also be analyzed in more concrete terms if we can better specify the structure of the problem.

As an illustration, though more general results are possible, we will concentrate on an optimal control problem of the type

$$(P) \quad \begin{cases} \text{Minimize in } u: & \displaystyle\int_0^T \left[\sum_{i=1}^n c_i(x(t)) u_i(t) + \sum_{i=1}^n c_{n+i}(x(t)) u_i^2(t) \right] dt \\ \quad\text{subject to} & x'(t) = Q_0(x(t)) + Q_1(x(t))u(t) + Q_2(x(t))u^2(t) \text{ in } (0,T), \\ & x(0) = x_0 \in R^n \text{ and } u(t) \in K \subset R^n. \end{cases}$$

We are taking here $N = n$. Q_1 and Q_2 are $n \times n$ matrices that, together with the vector Q_0, comply with appropriate technical hypotheses so that the state law is a well-posed problem. Set

$$Q = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix}, \quad c = \begin{pmatrix} c_1 & c_2 \end{pmatrix}, \tag{7}$$

where Q_1 is a nonsingular $n \times n$ matrix, and $c_1 \in R^n$. In addition, we put

$$\begin{gathered} D(x) = -(Q_1)^{-1} Q_2, \quad E(x) = c_1 D + c_2, \\ U(m, x) = 2 \sum_i m_i e_i \otimes e_i D - \mathrm{id}, \quad m = \phi(u), \end{gathered} \tag{8}$$

where the e_i's stand for the vectors of the canonical basis of R^n, and id is the identity matrix of size $n \times n$.

THEOREM 1.2. *Suppose that for the ingredients* (c, Q, K) *of* (P), *the following hold:*

1. *the matrix* U *is always nonsingular for* $u \in K$, *and* $x \in R^n$*;*
2. *for such pairs* (u, x), *we always have* $U^{-T} E < 0$, *componentwise.*

Then the optimal control problem admits solutions.

As a more specific example of the kind of existence result that can be obtained through this approach, we state the following corollary, whose proof amounts to going carefully through the arithmetic involved after Theorem 1.2.

Corollary 1.1. *Consider the optimal control problem*

$$\textit{Minimize in } u: \quad \int_0^T [c_1(x(t))(u_1(t))^2 + c_2(x(t))u_2(t)^2]\,dt$$

under

$$x_1'(t) = u_1(t) - u_2(t) + q_1(x)u_1(t)^2 + u_2(t)^2,$$
$$x_2'(t) = q_2(x)u_1(t) + u_2(t) + u_1(t)^2 + u_2(t)^2,$$

and an initial condition $x(0) = x_0$, where $u(t) \in K = [0,1]^2$,

$$q_1(x) \in \left(\frac{1}{3}, 1\right), \quad q_2(x) \in (-1,1), \quad c_1(x), c_2(x) > 0,$$

and

$$\frac{q_1(x)+1}{2}c_2(x) < c_1(x) < \frac{2(q_1(x))^2 + q_1(x)(q_2(x)+1) - q_2(x) - 3}{4(q_1(x)-1)}c_2(x).$$

Then there is at least one optimal solution of the problem.

Our strategy to prove these results is not new as it is based on the well-established philosophy of relying on relaxed versions of the original problem, and then, under suitable assumptions, proving that there are solutions of the relaxed problem which are indeed solutions of the original one (see [10], [15], [17], [18], [33], and [34]). From this perspective, it is a very good example of the power of relaxed versions in optimization problems.

The relaxed version of the problem that we will be using is formulated in terms of Young measures associated with sequences of admissible controls. These so-called parametrized measures were introduced by Young [34], [35], [36], [37], and have been extensively used in calculus of variations and optimal control theory (see, for example, [21], [22], [28], and [29]). Because of the special structure of the dependence on u, we will be concerned with (generalized) "moments" of such probability measures. Namely, the set

$$L = \{m \in R^s : m_i = \phi_i(u), 1 \le i \le s, u \in K\} \tag{9}$$

and the space of moments

$$\Lambda = \left\{m \in R^s : m_i = \int_K \phi_i(\lambda)\,d\mu(\lambda), 1 \le i \le s, \mu \in P(K)\right\} \tag{10}$$

will play a fundamental role. Here $P(K)$ is the convex set of all probability measures supported in K. Since the mapping

$$M : \mu \in P(K) \mapsto \Lambda, \quad M(\mu) = \int_K \phi(\lambda)\,d\mu(\lambda)$$

is linear, we easily conclude that Λ is a convex set of vectors and, in addition, that the set of its extreme points is contained in L. In fact, for some particular ϕ_i's of polynomial type, the set of the extreme points of Λ is precisely L. We examine and comment on the set Λ in section 3. This is closely related to the classical moment problem (see [1], [32] or more recently [11], [19]).

A crucial fact in our strategy is the following.

Assumption 1.1. For each fixed $x \in R^N$ and $\xi \in Q(x)\Lambda \subset R^N$, the minimum

$$\min_{m \in \Lambda} \{c(x) \cdot m : \xi = Q(x)m\}$$

is attained only in L.

It is interesting to realize the meaning of this assumption. If we drop the linear constraint $\xi = Qm$ on the above minimum, then the minimum is always attained in a certain point in L simply because a linear function on a convex set will always take its extreme values on extreme points of such a convex set. However, it is precisely the presence of the linear constraint $\xi = Qm$ that makes the hypothesis meaningful, as the extreme points of the section of Λ by such a set of linear constraints may not (indeed most of the time do not) belong to L, so that the extreme points of the linear function $c \cdot m$ over such a convex section may not attain its minimum on L. Our main hypothesis establishes that this should be so and, fundamentally, that the minimum is only attained in L.

Under this assumption and the other technical requirements indicated at the beginning, one can show a general existence theorem of optimal solutions for our problem.

Theorem 1.3. *Under Assumption* 1.1 *and the additional well-posedness hypotheses on* (c, Q) *indicated above, the initial optimal control problem* (P_1) *admits a solution.*

Notice that we are not assuming any convexity on the set K in this statement. The proof of this theorem can be found in section 2. As remarked before, the proof is more or less standard, and it involves the use of an appropriate relaxed formulation of the problem in terms of moments of Young measures [21], [29].

Condition (5) in Theorem 1.1 is nothing but a sufficient condition to ensure Assumption 1.1 in a more explicit way. As a matter of fact, all of our efforts are directed towards finding in various ways more explicit conditions for the validity of this assumption. In this vein, the rest of the paper focuses on exploring more fully our Assumption 1.1 either through duality or geometric arguments, or in order to prove Theorem 1.1. Ideally, one would like to provide explicit results saying that for a certain set $\mathcal{M}$, Assumption 1.1 holds if for each $x \in R^N$, $(c(x), Q(x)) \in \mathcal{M}$. In fact, by looking at Assumption 1.1 from the point of view of duality, one can write a general statement whose proof is a standard exercise.

Proposition 1.1. *If for any* $x \in R^N$, $(c, Q) = (c(x), Q(x))$ *are such that for every* $\eta \in R^N$ *there is a unique* $m(\eta) \in L$ *solution of the problem*

$$\text{(11)} \qquad \textit{Minimize in } m \in L: \quad (c + \eta Q)m,$$

then Assumption 1.1 *holds.*

We briefly comment on this in section 3. One then says that $(c, Q) \in \mathcal{M}$ if this pair verifies the condition on this proposition. A full analysis of this set $\mathcal{M}$ turns out to depend dramatically on the ingredients of the problem. In particular, we will treat the cases $n = N = 1$ and the typical situation of algebraic moments of degrees 2 and 3 in sections 4, 5, and 6. In section 7 we apply our results to a few explicit examples and compare it with the application of the classical Filippov–Roxin theory.

Situations where either $N > 1$ or $n > 1$ are much harder to deal with, especially because existence results are more demanding on the structure of the underlying problem. In particular, we need a convexity assumption on how the nonlinear dependence on controls occurs. We found that (5) turns out to be a general sufficient condition for the validity of Assumption 1.1, thus permitting us to prove Theorem 1.1 based on Theorem 1.3. Both Theorem 1.2 and Corollary 1.1 then follow directly from Theorem 1.1 after some algebra. This can be found in section 8.

Finally, we would like to point out that one particular interesting example, from the point of view of applications, that adapts to our results comes from the control of underwater vehicles (submarines). See [5], [14], and [16]. This served as a clear motivation for our work. We plan to go back to this problem in the near future.

2. Proof of Theorem 1.3. Consider the following four formulations of the same underlying optimal control problem.

(P_1) The original optimal control problem described in section 1.

(P_2) The relaxed formulation in terms of Young measures (see [21], [22], [28], [29]) associated with sequences of admissible controls:

$$\text{Minimize in } \mu = \{\mu_t\}_{t\in(0,T)} : \quad \tilde{I}(\mu) = \int_0^T \left[\int_K \sum_i c_i(x(t))\phi_i(\lambda)\, d\mu_t(\lambda) \right] dt$$

subject to

$$x'(t) = \int_K \sum_i Q_i(x(t))\phi_i(\lambda)\, d\mu_t(\lambda)$$

and

$$\operatorname{supp}(\mu_t) \subset K, \quad x(0) = x_0 \in R^N.$$

(P_3) The above relaxed formulation (P_2) rewritten by taking advantage of the moment structure of the cost density and the state equation. If we put $c = (c_1, \dots, c_s) \in R^s$, $Q \in M_{N\times s}$, and m such that

$$m_i = \int_K \phi_i(\lambda)\, d\mu_t(\lambda) \quad \forall i \in \{1, \dots, s\},$$

then we pretend to

$$\text{Minimize in } m \in \Lambda : \quad \int_0^T c(x(t)) \cdot m(t)\, dt$$

subject to

$$x'(t) = Q(x(t))m(t), \quad x(0) = x_0.$$

(P_4) Variational reformulation of formulation (P_3) (see [7], [23], [24], [27]). This amounts to defining an appropriate density by setting

$$\varphi(x, \xi) = \min_{m\in\Lambda}\{c(x) \cdot m \ : \xi = Q(x)m\}.$$

Then we would like to

$$\text{Minimize in } x(t) : \quad \int_0^T \varphi(x(t), x'(t))\, dt$$

subject to $x(t)$ being Lipschitz in $(0, T)$ and $x(0) = x_0$.

We know that the three versions of the problem, (P_2), (P_3), and (P_4), admit solutions because they are relaxations of the original problem (P_1). In fact, since K is compact, (P_2) is a particular case of the relaxed problems studied in [21] and [29]. The existence of solution for the linear optimal control problem (P_3) is part of the classical theory [7]. Indeed, (P_3) is nothing but (P_2) rewritten in terms of moments, so that the equivalence is immediate. (P_4) is the reformulated problem introduced in [27] whose equivalence to (P_3) was largely explored in [7] and [23], [24].

Let $\tilde{x}$ be one such solution of (P_4). By Assumption 1.1 applied to a.e. $t \in (0, T)$, we have

$$\varphi(\tilde{x}(t), \tilde{x}'(t)) = \min_{m \in \Lambda} \{c(\tilde{x}(t)) \cdot m(t) : \ \tilde{x}'(t) = Q(\tilde{x}(t))m(t)\} = c(\tilde{x}(t)) \cdot \tilde{m}(t)$$

for a measurable $\tilde{m}(t) \in L$, a solution of (P_3) (see [23]). The fundamental fact here (through Assumption 1.1) is that $\tilde{m}(t) \in L$ for a.e. $t \in (0, T)$, and this in turn implies that $\tilde{m}(t)$ is the vector of moments of an optimal Dirac-type Young measure $\mu = \{\mu_t\}_{t \in (0,T)} = \{\delta_{\tilde{u}(t)}\}_{t \in (0,T)}$ for an admissible $\tilde{u}$ for (P_1). This admissible control $\tilde{u}$ is optimal for (P_1). This finishes the proof.

3. The set Λ and duality.

The moment set Λ deserves some comments before proceeding further. Consider the mapping ϕ as in (1) and L as in (9).

We can regard L as part of an embedded n-manifold in R^s, $s > n$, and ϕ its standard or canonical parametrization. The moment set Λ defined in (10) is contained in the convex hull of this manifold.

The most important fact about Λ that one may need in our analysis is stated in the next proposition.

PROPOSITION 3.1. *The set of extreme points of* Λ *is contained in* L.

Proof. First notice, as shown in [19] in a context similar to ours, that the compactness of K implies

$$\overline{\text{co}(L)} = \text{co}(L) = \bar{\Lambda} = \Lambda.$$

The fact of K being bounded plays an important role because otherwise Λ can be shown to be not necessarily closed [11].

Since $\Lambda = \text{co}(L)$ then it is known from convex analysis [26] that

$$ext(\Lambda) \subseteq L,$$

where $ext(\Lambda)$ represents the extreme points of Λ. □

Remark 3.1. For some ϕ's it is possible to conclude that $ext(\Lambda) = L$. This is the case, for example, when ϕ contains all the linear and quadratic terms of an n-variable polynomial. However, this is not essential in what follows.

Due to this result the proof of Proposition 1.1 is standard (see [26]), so that we shall only make a few remarks.

Since

$$ext(\Lambda) \subseteq \text{co}(L),$$

which is a compact set, the minimum of

$$(c + \eta Q)m$$

in Λ is always attained at least in one point of L (it can be attained also in points of $\Lambda \setminus L$). However, if this point happens to be unique, because of Proposition 3.1, it is also immediate to check that it must be the unique minimizer in Λ.

The condition (11) in Proposition 1.1 means that

$$\min_{m\in\Lambda}(c+\eta Q)m = \min_{m\in L}(c+\eta Q)m = (c+\eta Q)\phi(a)$$

for a single $a \in K$, which also verifies

$$\min_{m\in\Lambda}(c+\eta Q)m - \eta\xi = (c+\eta Q)\phi(a) - \eta\xi$$

for $\xi \in Q(x)\Lambda$, that is, such that Assumption 1.1 is nonempty.

In particular the associated Karush–Kuhn–Tucker vector $\bar{\eta}$ verifies (see [26])

$$c\cdot\phi(a) + \bar{\eta}(Q\phi(a) - \xi) = \min_{m\in\Lambda}\{c\cdot m :\ Qm = \xi\} = c\cdot\phi(a)$$

for a single $a \in K$ complying with $Q\phi(a) = \xi$. As a consequence, for all admissible $m \in \Lambda$ different from $\phi(a)$, we have

$$c\cdot m > c\cdot\phi(a).$$

4. Polynomial dependence. The case $N = n = 1$, $p = 2$. Until section 8, we concentrate on the situation where

$$\phi : R^n \to R^s$$

is such that $\phi_i(u) = u_i$, $i = 1, 2, \ldots, n$, and $\phi_{n+i}(u)$, $i = 1, 2, \ldots, s-n$, are convex polynomials of some degree p, or simply polynomials whose restriction to K is convex. We will consider K itself to be convex.

Our goal is to explore different possibilities to apply Theorem 1.3 directly by ensuring Assumption 1.1. In other words, we will search for functions

$$c : R^N \to R^s, \quad Q : R^N \to R^{Ns},$$

such that for every $x \in R^N$,

$$(c(x), Q(x)) \in \mathcal{M},$$

where $\mathcal{M}$ represents the set

$$\left\{(c,Q) :\ \forall\ \xi \in Q\Lambda,\ \arg\min_{m\in\Lambda}\{c\cdot m : \xi = Qm\} \in L\right\}. \tag{12}$$

During the following three sections we will focus on the one-dimensional case $N = n = 1$ and use some ideas based in duality (Proposition 1.1) and in geometric interpretations.

In sections 4, 5, and 6, we explore various scenarios where Assumption 1.1 can be derived and defer explicit examples until section 7. In particular, we consider in this section the situation where ϕ is given by $\phi(a) = (a, a^2)$. We are talking about polynomial components of degree less than or equal to $p = 2$.

Let $K = [a_1, a_2]$, L, and Λ as in (9)–(10). The geometric interpretation of these sets is quite easy, as we can see in Figure 1. Here, we have $s = 2$, and

$$c : R \to R^2, \quad Q : R \to R^2$$

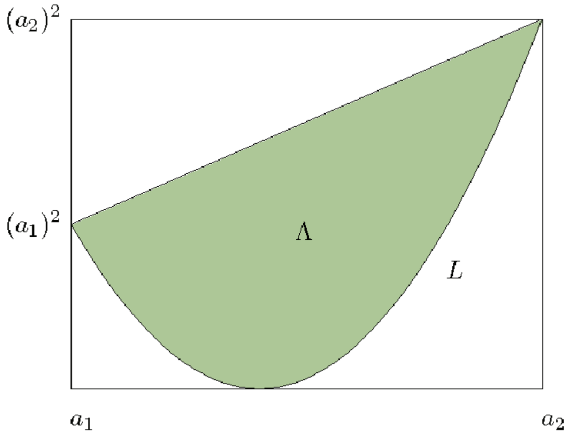


FIG. 1. $\Lambda = \mathrm{co}(L)$ *for* $p = 2$.

can be identified with vectors in R^2 or, more precisely, with plane curves parametrized by x. To emphasize that function Q is not matrix-valued but vector-valued, we will call it q.

Next we describe sufficient conditions for $(c(x), q(x)) \in \mathcal{M}$.

LEMMA 4.1. *Let K, L, and ϕ be as above. For every $x \in R$, let $q = q(x)$ and $c = c(x)$ be vectors such that one of the following conditions is verified:*

1. $q_1 + q_2(a_1 + a_2) = 0$ *and*
$$\det \begin{pmatrix} c_1 & c_2 \\ q_1 & q_2 \end{pmatrix} \neq 0;$$
2. $q_1 + q_2(a_1 + a_2) \neq 0$ *and*
$$(q_1 + q_2(a_1 + a_2)) \det \begin{pmatrix} c_1 & c_2 \\ q_1 & q_2 \end{pmatrix} < 0.$$

Then $(c, q) \in \mathcal{M}$, and consequently Assumption 1.1 *is verified.*

Proof. Suppose there is η such that the minimum of $(c + \eta q) \cdot m$ is attained in more than one point of $L = \phi(K)$. This means that the real function

$$g(t) = (c + \eta q) \cdot \phi(t) = (c_1 + \eta q_1)t + (c_2 + \eta q_2)t^2$$

has more than one minimum point over K. For that to happen, either g is constant on t, i.e.,

$$\begin{cases} c_1 + \eta q_1 = 0 \\ c_2 + \eta q_2 = 0 \end{cases} \Leftrightarrow \det \begin{pmatrix} c_1 & c_2 \\ q_1 & q_2 \end{pmatrix} = 0,$$

which contradicts our hypothesis, or else we must have

$$c_2 + \eta q_2 < 0, \quad g'\left(\frac{a_1 + a_2}{2}\right) = 0.$$

This condition can be written as

$$c_1 + (a_1 + a_2)c_2 + \eta[q_1 + (a_1 + a_2)q_2] = 0.$$

If $q_1 + q_2(a_1 + a_2) = 0$, but $c_1 + (a_1 + a_2)c_2 \neq 0$ (condition 1 of Lemma 4.1), then this equation can never be fulfilled. Otherwise, there is a unique value for η, by solving this equation, which should also verify the condition on the sign of $c_2 + \eta q_2$. It is elementary, after going through the algebra, that the condition on this sign cannot be true under the second condition on the statement of the lemma. □

5. The case $N = n = 1$, $p = 3$. We study the case where $\phi(a) = (a, a^2, a^3)$, $s = 3$, and c and q can be identified as vectors in R^3. The understanding of the set Λ and its sections by planes in R^3 is much more subtle, however. In Figure 2 we can see Λ and L for $a_1 > 0$.

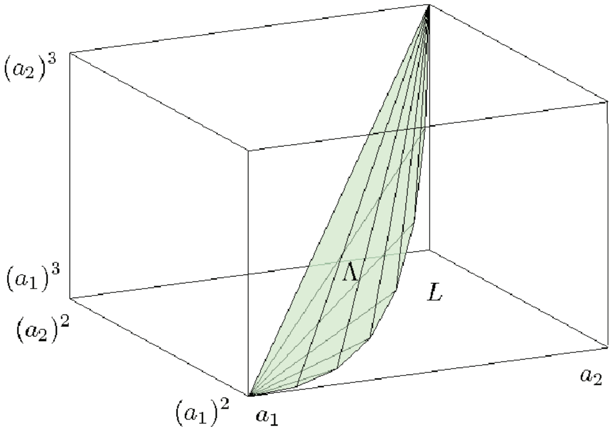


FIG. 2. $\Lambda = \mathrm{co}(L)$ *for* $p = 3$.

To repeat the procedure used for $p = 2$, and apply Proposition 1.1, we would like to give sufficient conditions for the function

$$g(t) = (c + \eta q) \cdot \phi(t) = (c_1 + \eta q_1)t + (c_2 + \eta q_2)t^2 + (c_3 + \eta q_3)t^3 \tag{13}$$

to have a single minimum over $K = [a_1, a_2]$ for every η. As indicated, and after some reflection, a complete analysis of the situation is rather confusing and the conditions on vectors c and q much more involved. To illustrate this, we give a sufficient condition in the following form.

LEMMA 5.1. *For all $x \in R$, let $c = c(x)$ and $q = q(x)$ be vectors in R^3 such that*

$$q_2^2 - 3q_1q_3 < 0, \quad (2c_2q_2 - 3c_1q_3 - 3q_1c_3)^2 - 4(c_2^2 - 3c_1c_3)(q_2^2 - 3q_1q_3) < 0;$$

then $(c, q) \in \mathcal{M}$, and Assumption 1.1 *is verified.*

Proof. The proof consists in the realization that the conditions on vectors c and q ensure that the cubic polynomial (13) is monotone in all of R (avoiding degenerate situations), and thus it can attain the minimum only in a single point of any finite interval. Notice that this condition is independent of the interval. In fact, we have to discard the possibility for the derivative of the polynomial $g(t)$ to have roots. This amounts to the negativity of the corresponding discriminant. And this, in turn, is a quadratic expression in η that ought to always be negative. This occurs when that parabola has a negative discriminant, and the leading coefficient is also negative. These two conditions are exactly those in the statement of this lemma. □

A more general condition would focus on considering the local maximizer and the local minimizer of $g(t)$, M_+ and M_-, respectively, and demanding that the interval $[a_1, a_2]$ have an empty intersection with the interval determined by M_+ and M_-. But this would lead to rather complicated expressions. Even so, sometimes under more specific hypotheses on the form of vectors c and q these conditions can be exploited.

Remark 5.1. Notice that the relation

$$ext(\Lambda) = L$$

is not true for a general K if it has positive and negative values. However, it is true if we consider $a_1 > 0$ or $a_2 < 0$.

LEMMA 5.2. *Let $K = [a_1, a_2]$ with $a_1 > 0$ and*

$$(c, q) = ((0, c_2, c_3), (0, q_2, q_3))$$

such that

$$-\frac{q_2}{q_3} < 0, \quad (c_2, c_3) \cdot \left(1, -\frac{q_2}{q_3}\right) < 0.$$

Then the assumptions of Proposition 1.1 *are valid, and consequently so is Assumption* 1.1*.*

Proof. In this situation, the maximizer M_+ referred to above is given by

$$M_+ = \frac{-(c_2 + \eta q_2) - |c_2 + \eta q_2|}{3(c_3 + \eta q_3)}.$$

If $q_2 > 0$, then $q_3 > 0$, and we have

$$\frac{c_2}{q_2} > -\frac{c_3}{q_3}.$$

Hence if $\eta \in \left]-\infty, -\frac{c_2}{q_2}\right] \setminus \{-\frac{c_3}{q_3}\}$, then

$$M_+ = \frac{-(c_2 + \eta q_2) + c_2 + \eta q_2}{3(c_3 + \eta q_3)} = 0.$$

If $\eta > -\frac{c_2}{q_2}$, then

$$M_+(\eta) = \frac{-(c_2 + \eta q_2) - (c_2 + \eta q_2)}{3(c_2 + \eta q_2)} = \frac{-2(c_2 + \eta q_2)}{3(c_3 + \eta q_3)} < 0.$$

In any case $M_+(\eta) \leq 0$, thus $a_1 > M_+$.

Also if $\eta = -\frac{c_3}{q_3}$, then

$$g(t) = (c_2 + \eta q_2)t^2 = (c_2, c_3) \cdot \left(1, -\frac{q_2}{q_3}\right) t^2,$$

which has a unique minimum in K since we have assumed $a_1 > 0$. We conclude that the condition (11) in Proposition 1.1 is verified. ▯

In a very similar way we can prove the following.

LEMMA 5.3. *Let $K = [a_1, a_2]$ with $a_2 < 0$ and*

$$(c, q) = ((0, c_2, c_3), (0, q_2, q_3))$$

such that

$$-\frac{q_2}{q_3} > 0, \quad (c_2, c_3) \cdot \left(1, -\frac{q_2}{q_3}\right) < 0.$$

Then $(c, q) \in \mathcal{M}$, and consequently Assumption 1.1 *is valid.*

6. A geometric approach to the case $N = n = 1$, $p = 3$. As we have seen, the use of Proposition 1.1 is simpler only when restricted to some particular classes of examples. Thus we propose a general criteria for obtaining Assumption 1.1 based on a geometric approach.

We first give a result that generalizes the strict convexity of a ϕ-parametrized plane curve for a three-dimensional one.

LEMMA 6.1. *Let $K = [a_1, a_2]$ with $a_1 > 0$, $\phi(t) = (t, t^2, t^3)$, and L the curve parametrized by ϕ for t in K.*

1. *Given t in K, then for all $s \in K$ such that $s \neq t$, we have*

$$(\phi(s) - \phi(t)) \cdot N(t) > 0,$$

where $N(t)$ is the normal vector to ϕ at t.

2. *For every $t \in K$, $v \in \Lambda = \mathrm{co}(L) \setminus \{\phi(t)\}$, we have*

$$(v - \phi(t)) \cdot N(t) > 0.$$

Proof. To check the first part of the statement notice that since

$$\phi'(t) = (1, 2t, 3t^2)$$

and

$$\phi''(t) = (0, 2, 6t)$$

we have that the normal vector, colinear to $\phi'(t) \times \phi''(t)$, is given by

$$N(t) = c_t(-9t^2 - 2t, 1 - 9t^4, 6t^3 + 3t),$$

where $c_t > 0$ is a normalizing constant. Setting

$$N_1 = -9t^3 - 2t, \quad N_2 = 1 - 9t^4, \quad N_3 = 6t^3 + 3t,$$

we find that the solution s of

$$(\phi(s) - \phi(t)) \cdot N(t) = 0$$

also verifies

$$N_3 s^3 + N_2 s^2 + N_1 s - N \cdot \phi(t) = 0,$$

which is equivalent to

$$(s - t)^2 (N_3 s + N_2 + 2t N_3) = 0.$$

This means that

$$\hat{s} = -\frac{N_2}{N_3} - 2t = -\frac{3t^4 + 6t^2 + 1}{6t^3 + 3t}$$

is the only solution different from t, but also that it is negative for all $t > 0$, and consequently that it should be excluded. Once we assumed $K \subset R^+$ and $s \neq t$ the conclusion is immediate.

By using the previous discussion, proving the second part of the statement is trivial once we notice that both m and $\phi(t)$ can be rewritten as

$$\sum_{i=1}^{4} \alpha_i \phi(s_i) \quad \text{and} \quad \sum_{i=1}^{4} \alpha_i \phi(t),$$

respectively, where $s_i \in K$ and $\sum_{i=1}^{4} \alpha_i = 1$. ◻

Another useful lemma follows.

Lemma 6.2. *If q and c are such that*

$$(\phi'(t) \times (c \times q)) \cdot (\phi(s) - \phi(t)) \tag{14}$$

does not change sign for $t, s \in K$, $s \neq t$, then if $v \in \Lambda$, $v \neq \phi(t)$, and $q \cdot (v - \phi(t)) = 0$, we have

$$c \cdot (v - \phi(t)) \neq 0.$$

This means that the linear function c cannot take the same value over $\phi(t)$ and any $v \neq \phi(t)$ in the plane section

$$\{v \in \Lambda : \ q \cdot v = q \cdot \phi(t)\}.$$

Proof. Notice that for $v \in \Lambda$,

$$\begin{aligned}(\phi'(t) \times (c \times q)) \cdot (v - \phi(t)) &= (\phi'(t) \times (c \times q)) \cdot \left(\sum_{i=1}^{4} \alpha_i \phi(s_i) - \sum_{i=1}^{4} \alpha_i \phi(t) \right) \\ &= \sum_{i=1}^{4} \alpha_i (\phi'(t) \times (c \times q)) \cdot (\phi(s_i) - \phi(t)) > 0 \ (\text{or } < 0),\end{aligned}$$

so that the condition stated is also verified for any $v \in \Lambda$.

Suppose now that $v \in \Lambda$ verifies $q \cdot (v - \phi(t)) = 0$ for given $t \in K$ with $v \neq \phi(t)$ and is such that $c \cdot (v - \phi(t)) = 0$. Then

$$\begin{aligned}(\phi'(t) \times (c \times q)) \cdot (v - \phi(t)) &= [(\phi'(t) \cdot q)c - (\phi'(t) \cdot c)q] \cdot (v - \phi(t)) \\ &= (\phi'(t) \cdot q)c \cdot (v - \phi(t)) - (\phi'(t) \cdot c)q \cdot (v - \phi(t)) = 0,\end{aligned}$$

a contradiction concerning the argument above. ◻

We now define the set $\mathcal{M}_1$ of pairs $(c, q) \in R^3 \times R^3$ through the following requirements:

(i) the quantity in (14) does not change sign over the pairs $t, s \in K$, $s \neq t$;

(ii) whenever there is a unique $a \in K = [a_1, a_2]$ such that

$$(\phi(a_1) + \phi(a_2) - 2\phi(a)) \cdot q = 0, \tag{15}$$

then

$$(\phi(a_1) + \phi(a_2) - 2\phi(a)) \cdot c > 0.$$

Once more we can establish the following result.

Proposition 6.1. *Let $\mathcal{M}$ be as in* (12).

If $a_1 > 0$ and $(c, q) \in \mathcal{M}_1$, then $(c, q) \in \mathcal{M}$ and Assumption 1.1 *holds.*

Proof. 1. Suppose first that there is $a \in K$ such that we have (15).

Let

$$v_a = \frac{\phi(a_1) + \phi(a_2)}{2}.$$

Consider $v \in \Lambda$ such that

$$[v - \phi(a)] \cdot q = 0.$$

Suppose

$$c \cdot [v - \phi(a)] < 0,$$

and consider the continuous function

$$G(v, u) = c \cdot (v - u)$$

over the bounded path connecting $(v_a, \phi(a))$ and $(v, \phi(a))$ given by

$$S = \{\alpha[(v, \phi(a)) - (v_a, \phi(a))] + (v_a, \phi(a)) : \ \alpha \in [0, 1]\}.$$

It is easy to check that every component of a vector of S is contained in the section

$$\{v \in \Lambda : q \cdot v = q \cdot \phi(a)\}.$$

Then there exists α such that

$$G(\alpha[(v, \phi(a)) - (v_a, \phi(a))] + (v_a, \phi(a))) = 0,$$

or in other words

$$c \cdot [\alpha(v - v_a) + v_a - \phi(a)] = 0,$$

which by Lemma 6.2 means that necessarily

$$\alpha(v - v_a) + v_a = \phi(a).$$

Consequently

$$\alpha(v - \phi(t)) \cdot N(t) + (1 - \alpha)(v_t - \phi(t)) \cdot N(t) = 0,$$

and this is in contradiction with Lemma 6.1. Hence

$$c \cdot [v - \phi(a)] > 0 \text{ if } c \cdot [v_a - \phi(a)] > 0.$$

Let $\bar{t}$ be such that

$$q \cdot \phi(a_1) = q \cdot \phi(\bar{t})$$

and $t \neq a$, $t \geq \bar{t}$, such that

$$v_t = \alpha[\phi(a_2) - \phi(a_1)] + \phi(a_1) \in \Lambda$$

verifies

$$[v_t - \phi(t)] \cdot q = 0.$$

Considering once more the continuous function $G(v,u)$ over the path connecting $(v_t,\phi(t))$ and $(v_a,\phi(a))$, as

$$\alpha[v_t - v_a] + v_a \in \{v : q \cdot v = q \cdot \phi(t)\},$$

we can, as we did above, conclude that if

$$c \cdot [v_t - \phi(t)] < 0,$$

then for certain α,

$$\alpha[v_t - v_a] + v_a = \phi$$

and, consequently,

$$c \cdot [v_t - \phi(t)] > 0$$

for any $t \geq \bar{t}$. The same type of argument shows that

$$c \cdot [v - \phi(t)] > 0$$

for any v such that

$$q \cdot v = q \cdot \phi(t).$$

If $t < \bar{t}$, there exists $s \in K$ such that

$$q \cdot \phi(s) = q \cdot \phi(t).$$

In this situation, again the continuity of G should be applied to the path connecting

$$(v_{\bar{t}}, \phi(\bar{t})) = (\phi(a_1), \phi(\bar{t}))$$

and

$$(\phi(s), \phi(\bar{t})),$$

repeatedly until the limit case when $\phi(s) = \phi(\bar{t})$.

If there is $\bar{t} \neq a_2$ such that

$$q \cdot \phi(a_2) = q \cdot \phi(\bar{t}),$$

we shall proceed in an analogous way.

2. Suppose now that there are $a, b \in K$ such that

$$(v_a - \phi(a)) \cdot q = (v_a - \phi(b)) \cdot q = 0.$$

Then it is not difficult to conclude that

$$a = a_1 \quad \text{and} \quad b = a_2.$$

Hence assuming (without loss of generality) that

$$(\phi(a_1) - \phi(a_2)) \cdot c > 0$$

we can, once again, use the continuity of G to conclude

$$c \cdot [\phi(s) - \phi(t)] > 0,$$

where $\phi(s)$ and $\phi(t)$ verify

$$(\phi(s) - \phi(t)) \cdot q = 0$$

and, after that, for a general v such that

$$(v - \phi(b)) \cdot q = 0. \qquad \square$$

Remark 6.1. This type of argument can be also deduced for the case $N = n = 1$, $p = 2$, where it can be seen to be equivalent to the conditions in Lemma 4.1. However, when the parameters N, n, and p increase their values, it becomes very hard to give geometrically based sufficient conditions in such an exhaustive manner as we have done here. Even so, in section 8 we show how to give more restrictive yet more general sufficient conditions (Theorems 1.2 and 1.1) for interesting high-dimensional particular situations, where some geometrical ideas can be used as a way to verify Assumption 1.1.

7. Examples.

Before going further to higher-dimensional situations we gather in this section some typical, academic examples for which either Lemma 4.1, Lemma 5.2, or Proposition 6.1 can be applied.

7.1. Example 1.

Let us consider the optimal control problem

$$\text{Minimize in } u: \quad \int_0^T [c(x(t))u(t) + u^2(t)]\, dt$$

under

$$x'(t) = q(x(t))u(t) + u^2(t), \quad x(0) = x_0,$$

where $|u(t)| \le 1$.

We have the following remarkable existence result.

Lemma 7.1. *If the functions q and c are Lipschitz, and*

$$q(q - c) > 0,$$

then the optimal control problem admits a solution.

The proof reduces to performing some elementary algebra to check the conditions of Lemma 4.1.

Instead of applying that lemma, as both our cost and state-equation functions have cross dependence on x and on u so that we can't apply the results of [3], [4], [25], one can try the classical existence result based on the classical Filippov–Roxin theory. For that we need to check if the orientor field

$$\mathcal{A}_x = \{(\xi, v) : \ v \ge c(x)u + u^2, \ \xi = q(x)u + u^2, \ u \in K = [-1, 1]\}$$

is a convex set. Notice that K is bounded so coercivity is not an issue here. Proceeding in that direction, we can see that

$$\xi = q(x)u + u^2$$

is equivalent to

$$u_1 = -\frac{q + \sqrt{q^2 + 4\xi}}{2} \quad \text{or} \quad u_2 = -\frac{q - \sqrt{q^2 + 4\xi}}{2},$$

which are possible solutions when ξ is such that $\xi \geq -\frac{q^2}{4}$ and at least one of them belongs to $K = [-1, 1]$. Letting

$$F_i(x, \xi) = c(x)u_i + u_i^2, \ i = 1, 2,$$

we see that

$$F_2 \leq F_1$$

for all ξ as above. Consequently

$$\begin{aligned} \mathcal{A}_x &= \mathcal{A}_x^1 \cup \mathcal{A}_x^2 \\ &= \left\{ (\xi, v) : \ v \geq F_2(x, \xi), \xi \in u_2^{-1}(K) \cap \left[-\frac{q^2}{4}, +\infty \right[\right\} \\ &\cup \left\{ (\xi, v) : \ v \geq F_1(x, \xi), \xi \in \left(u_1^{-1}(K) \setminus u_2^{-1}(K) \right) \cap \left[-\frac{q^2}{4}, +\infty \right[\right\}, \end{aligned}$$

where, for $i = 1, 2$, u_i^{-1} refers to the preimage of the solutions u_i as functions of ξ.

Because of the assumption on (c, q) it is easy to see that $\mathcal{A}_x^2 = \emptyset$, and consequently that the convexity of $\mathcal{A}_x$ reduces to the convexity of the function

$$F_2(\xi) = \frac{q - c}{2}(q - \sqrt{q^2 + 4\xi})$$

over a certain convex set

$$u_2^{-1}(K) \cap \left[-\frac{q^2}{4}, +\infty \right[.$$

This can be checked by elementary calculus.

We now turn to the possibility of applying the result in [9] to this example. First, in order to write our problem as a minimum-time problem, we require that $c(x)u + u^2$ never change sign in $R \times K$ [7]. So a first restriction must be imposed. For example, consider $c(.)$ and $q(.)$ such that

$$q(x) > c(x) > 1.$$

The right member of the differential equation of the minimum-time problem is given by

$$\mathcal{F}(x, K) = \left\{ \frac{q(x)u + u^2}{c(x)u + u^2} : \ u \in K \right\}.$$

The result in [9] doesn't ask for the convexity of the set-valued map $\mathcal{F}$, but it requires a linear boundedness in the sense that

$$\exists \alpha, \ \beta \text{ s.t. } \forall x \in R, \ \forall \xi \in \mathcal{F}(x, K) \text{ then}$$
$$\|\xi\| \leq \alpha \|x\| + \beta.$$

It is easy to see that this condition places a real constraint on the relative growth of pairs (c, q), even before verifying the remaining assumptions in [9].

7.2. Example 2. Look at the problem

$$\text{Minimize in } u: \quad \int_0^T [c(x(t))u^2(t) + u^3(t)]\, dt$$

under

$$x'(t) = [q(x(t))]u^2(t) + u^3(t), \quad x(0) = x_0,$$

where $u(t) \in [a_0, a_1]$, $a_0 > 0$.

LEMMA 7.2. *If the functions $q(x)$ and $c(x)$ are Lipschitz,*

$$c(x) < q(x) \ \forall x,$$

and $q(x)$ is always positive, then the optimal control problem admits solutions.

This result comes directly by applying Lemma 5.2 and Theorem 1.3.

Let us see what we would need to do if, alternatively, we decided to use the classical existence theory.

As we have seen in the previous example, we need to check the convexity of the orientor field

$$\mathcal{A}_x = \{(\xi, v) : \ v \geq c(x)u^2 + u^3, \ \xi = q(x)u^2 + u^3, \ u \in K = [a_0, a_1]\}.$$

In this case, according to the discriminant

$$\Delta = 27\xi^2 - 4\xi q$$

of the equation

$$\xi = q(x)u^2 + u^3,$$

we will have from one to three possible real solutions. Consider for each ξ

$$F_i = c(x)u_i^2 + u_i^3,$$

such that

$$F_1 \leq F_2 \leq F_3,$$

where $u_i = u_i^x(\xi), \ i = 1, 2, 3$, are the three, possibly equal, real solutions. Then

$$\begin{aligned} \mathcal{A}_x &= \mathcal{A}_x^1 \cup \mathcal{A}_x^2 \cup \mathcal{A}_x^3 \\ &= \{(\xi, v) : \ v \geq F_1, \ \xi \in u_1^{-1}(K)\} \\ &\cup \{(\xi, v) : \ v \geq F_2, \ \xi \in u_2^{-1}(K) \setminus u_1^{-1}(K)\} \\ &\cup \{(\xi, v) : \ v \geq F_3, \ \xi \in u_3^{-1}(K) \setminus \big(u_2^{-1}(K) \cup u_1^{-1}(K)\big)\}. \end{aligned}$$

Checking the convexity of this set or, alternatively, of the function

$$\varphi_x(\xi) = \begin{cases} F_1(\xi) & \text{if} \quad \xi \in u_1^{-1}(K), \\ F_2(\xi) & \text{if} \quad \xi \in u_2^{-1}(K) \setminus u_1^{-1}(K), \\ F_3(\xi) & \text{if} \quad \xi \in u_3^{-1}(K) \setminus \big(u_2^{-1}(K) \cup u_1^{-1}(K)\big) \end{cases}$$

is not an easy task at all, especially when compared to the almost immediate exercise of verifying the conditions of Lemma 5.2. It is also plausible that the inherent difficulties in applying classical theory will increase until a practically impossible scenario in which we let N, n, and p grow.

7.3. Example 3. In order to give a heuristic for using the criteria given in Proposition 6.1, let us consider the previous problem just by rewriting q as $c-\beta$ and for a specific K.

$$\text{Minimize in } u: \quad \int_0^T [c(x(t))u^2(t)+u^3(t)]\,dt$$

under

$$x'(t)=[c(x(t))-\beta(x(t))]u^2(t)+u^3(t), \quad x(0)=x_0,$$

where $u(t)\in[1,2]$.

LEMMA 7.3. *If the functions β and c are Lipschitz, and*

$$\beta<\min\{0,c\},$$

then the optimal control problem admits solutions.

Proof. First notice that for $a\in K=[a_1,a_2]$, we can find α such that the vector

$$B=\alpha[\phi(a_2)-\phi(a_1)]+\phi(a_1)$$

verifies

$$[B-\phi(a)]\cdot q=0.$$

Moreover, it is not difficult to see that

$$\alpha=\frac{a^3-a_1^3-m(a^2-a_1^2)}{a_2^3-a_1^3-m(a_2^2-a_1^2)},$$

and in the projection plane yz, (B_2,B_3) belongs to the line of slope m passing through (a^2,a^3),

$$B_3-a^3=m(B_2-a^2),$$

where

$$B_2-a^2=\frac{(a-a_1)[a^2(a_2+a_1)-a^2(a+a_1)]}{a_2^2+a_1a_2+a_1^2-m(a_2+a_1)}$$

and $m=-\frac{q_2}{q_3}$.

In our case $K=[1,2]$, so because of what we have just seen, taking $a_1=1$ and $a_2=2$ we see that for $a\in K$, we can find

$$\alpha=\frac{a^3-ma^2+m-1}{7-3m}\in[0,1]$$

such that

$$[\alpha[\phi(a_2)-\phi(a_1)]+\phi(a_1)-\phi(a)]\cdot q=0,$$

where

$$m=-\frac{c-\beta}{1}=\beta-c<0.$$

Furthermore, it is easy to see that the equation $\alpha = \frac{1}{2}$ has a unique solution in K. Consequently, if we consider $q = (0, c-\beta, 1)$ and $\bar{c} = (0, c, 1)$, there exists a unique $a \in K$ such that

$$[\phi(1) - \phi(0) - 2\phi(a)] \cdot q = 0.$$

Also, because of what we have seen above,

$$\left[\frac{1}{2}(\phi(1) - \phi(0)) - \phi(a)\right] \cdot \bar{c} = (B_2 - a^2)c + (B_3 - a^3) = (B_2 - a^2)(c + m)$$

$$= \frac{(a-1)(3a^2 - 4a - 4)}{7 - 3m}(c + \beta - c) > 0.$$

In addition, given $t, s \in K,\ s \neq t$,

$$(\phi'(t) \times (c \times q)) \cdot (\phi(s) - \phi(t)) = 0 \Leftrightarrow$$

$$\beta t(0, 3t, -2) \cdot (\phi(s) - \phi(t)) = 0 \Leftrightarrow$$

$$(s - t)[3t(s + t) - 2(s^2 + st + t^2)] = 0 \Leftrightarrow$$

$$s = -\frac{t}{2} \vee s = t,$$

which is impossible since $s \in K = [1, 2]$ and $s \neq t$. The result follows then by applying Proposition 6.1. □

8. The case N, $n > 1$. The previous analysis makes it very clear that checking Assumption 1.1 may be a very hard task as soon as n and/or N become greater than 1. Yet in this section we would like to show that there are chances to prove some nontrivial results.

The three main ingredients in Assumption 1.1 are

(i) the vector $c \in R^s$ in the cost functional;
(ii) the matrix $Q \in M_{N\times s}$ occurring in the state equation;
(iii) the convexification Λ of the set of moments L.

For (c, Q) given, consider the set $\mathcal{N}(c, Q)$ as it was defined in (3). Let Ψ be as in (2) and such that $\nabla\Psi(m)$ is a rank $s - n$ matrix and L can be seen as the embedded (parametrized) manifold of R^s in the manifold defined implicitly by $\Psi = 0$. This means that $\Psi(\phi(u)) = 0$ for all $u \in K$.

Consider also the set of vectors $\mathcal{N}(K, \phi)$ described in (4), that is, the set of "ascent" directions for Ψ at points of L.

We are now in a position to prove Theorem 1.1.

Proof. The proof is rather straightforward. First, note that due to the convexity assumption on Ψ, and the fact that $L \subset \{\Psi = 0\}$, we have $\Lambda \subset \{\Psi \leq 0\}$.

Suppose that $m_0 \in L$ and $m_1 \in \Lambda$, so that

$$\Psi(m_0) = 0,\ \Psi(m_1) \leq 0,\ cm_1 \leq cm_0,\ \text{and}\ Qm_1 = Qm_0\ (= \xi).$$

Then it is obvious that $m = m_1 - m_0 \in \mathcal{N}(c, Q)$. Because of our assumption, $m \in \mathcal{N}(K, \phi)$. We have two possibilities:

1. $\nabla\Psi(m_0)m = 0$. Because of the convexity of each component of Ψ, we have

$$\Psi(m_1) - \Psi(m_0) - \nabla\Psi(m_0)m \geq 0.$$

But then

$$0 = \Psi(m_0) \le \Psi(m_1) \le 0,$$

so that $m_1 \in L$. Because of the strict convexity of each component of Ψ, this means that $m_1 = m_0$, and Assumption 1.1 holds.

2. $\nabla\psi_i(m_0)m > 0$ for some i. Once again we have

$$\psi_i(m_1) - \psi_i(m_0) - \nabla\psi_i(m_0)m \ge 0.$$

But this is impossible because $\psi_i(m_1) > 0$ cannot happen for a vector in Λ. □

Remark 8.1. Notice that if in the original problem (P_1) we would have considered the dynamics given by

$$Q(x)\phi(u) + Q_0(x)$$

instead of just $Q(x)$, Assumption 1.1 and Theorem 1.1 could be written in exactly the same way.

Though Theorem 1.1 can be applied to more general cases, we will focus on a particular situation motivated by the control of underwater vehicles [5]. We will briefly describe the structure of the state equation. Indeed, it is just

$$x'(t) = Q_1(x)\phi(u) + Q_0(x),$$

where the state $x \in R^{12}$ incorporates the position and orientation in body and world coordinates, and the control $u \in R^{10}$ accounts for guidance and propulsion. Under suitable simplifying assumptions (see [5], [14], [16]), the components of the control vector u occur only as either linear or pure squares, in such a way that $\phi(u) = (u, u^2) \in R^{20}$ and $u^2 = (u_i^2)_i$, componentwise. Q_1 and Q_0 are matrices which may have essentially any kind of dependence on the state x.

To cover the sort of situations just described, we will concentrate on the optimal control problem (P) already stated in section 1 and set D, E, and U as in (7)–(8).

We can now prove Theorem 1.2.

Proof of Theorem 1.2. Notice that accordingly to (6), as $s = 2n$, we have, for $m \in R^s$,

$$\psi_i(m) = m_i^2 - m_{n+i}, \quad i = 1, 2, \dots, n,$$

which are certainly smooth and (strictly) convex. Moreover,

$$\nabla\Psi(m) = \begin{pmatrix} 2\tilde{m} & -\mathrm{id} \end{pmatrix},$$

where

$$\tilde{m} = 2\sum_i m_i e_i \otimes e_i,$$

and e_i is the canonical basis of R^n.

Suppose we have, for a vector $v \in R^{2n}$, $v = (v_1, v_2)$, that

$$Qv = 0, \quad cv \le 0.$$

A more explicit way of writing this is

$$Q_1 v_1 + Q_2 v_2 = 0, \quad c_1 v_1 + c_2 v_2 \le 0.$$

So

$$v_1 = Dv_2, \quad Ev_2 \leq 0.$$

We have to check that such a vector v is not a direction of descent for every function ψ_j, or it is an ascent direction for at least one of them. Note that

$$\nabla\Psi(m)v = Uv_2, \quad Ev_2 \leq 0.$$

It is an elementary linear algebra exercise to check that if $U^{-T}E < 0$, then condition (5) is verified so that Theorem 1.1 can be applied. □

Corollary 1.1 is a specific example of the kind of existence result that can be obtained through this approach. Its proof amounts to going carefully through the arithmetic while checking that matrix U and vector E defined from such a given class of $(c(.), Q(.))$ verify the assumptions of Theorem 1.2.

By using the same ideas, more general situations can be treated; for example, the number of controls could be greater than the components of the state. This is in fact the situation in the model that has served as an inspiration for us. We will pursue a closer analysis of such a particular situation, even stressing the more practical issues, in a forthcoming work.

REFERENCES

[1] N. I. Akhiezer, *The Classical Moment Problem and Some Related Questions in Analysis*, Oliver and Boyd, Edinburgh, London, 1961 (translated from Russian).

[2] E. J. Balder, *New existence results for optimal controls in the absence of convexity: The importance of extremality*, SIAM J. Control Optim., 32 (1994), pp. 890–916.

[3] H. Bauer, *Minimalstellen von funktionen und extremalpunkte*, Arch. Math., 9 (1958), pp. 389–393.

[4] H. Bauer, *Minimalstellen von funktionen und extremalpunkte* II, Arch. Math., 9 (1958), 11 (1960), pp. 200–205.

[5] D. Brutzman, *A Virtual World for an Autonomous Underwater Vehicle*, Dissertation, Naval Postgraduate School, Monterey, CA, 1994.

[6] L. Cesari, *An existence theorem without convexity conditions*, SIAM J. Control, 12 (1974), pp. 319–331.

[7] L. Cesari, *Optimization Theory and Applications: Problems with Ordinary Differential Equations*, Springer-Verlag, Berlin, 1983.

[8] A. Cellina and G. Colombo, *On a classical problem of the calculus of variations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 7 (1990), pp. 97–106.

[9] A. Cellina, A. Ferriero, and E. M. Marchini, *On the existence of solutions to a class of minimum time control problems and applications to Fermat's principle and to the brachystocrone*, Systems Control Lett., 55 (2006), pp. 119–123.

[10] F. Clarke, *Admissible relaxation in variational and control problems*, J. Math. Anal. Appl., 51 (1975), pp. 557–576.

[11] J. J. Egozcue, R. Meziat, and P. Pedregal, *From a nonlinear, nonconvex variational problem to a linear, convex formulation*, Appl. Math. Optim., 47 (2003), pp. 27–44.

[12] I. Ekeland and R. Temam, *Analyse convexe et problemes variationnels*, Dunod, Paris, 1974.

[13] A. F. Filippov, *On certain questions in the theory of optimal control*, SIAM J. Control, 1 (1962), pp. 76–84.

[14] T. Fossen, *Guidance and Control of Ocean Vehicles*, John Wiley & Sons, Chichester, UK, 1994.

[15] R. V. Gamkrelidze, *On sliding optimal states*, Dokl. Akad. Nauk SSSR, 143 (1962), pp. 1243–1245 (in Russian).

[16] A. J. Healey and D. Lienard, *Multivariable sliding mode control for autonomous diving and steering of unmanned underwater vehicles*, IEEE J. Oceanic Engrg., 18 (1993), pp. 327–339.

[17] E. J. McShane, *Curve-space topologies associated with variational problems*, Ann. Scuola Norm. Super. Pisa (2), 9 (1940), pp. 45–60.

[18] E. J. McShane, *Relaxed controls and variational problems*, SIAM J. Control, 5 (1967), pp. 438–485.

[19] R. Meziat, *Analysis of non convex polynomial programs by the method of moments*, in Frontiers in Global Optimization, Nonconvex Optim. Appl. 74, Kluwer Academic, Boston, MA, 2004, pp. 353–371.

[20] B. S. Mordukhovich, *Existence theorems in nonconvex optimal control*, in Calculus of Variations and Optimal Control, A. Ioffe, S. Reich, and I. Shafrir, eds., Chapman & Hall/CRC Press, Boca Raton, FL, 1999, pp. 173–197.

[21] J. Munoz and P. Pedregal, *A refinement on existence results in nonconvex optimal control*, Nonlinear Anal., 46 (2001), pp. 381–398.

[22] P. Pedregal, *Parametrized Measures and Variational Principles*, Progr. Nonlinear Differential Equations Appl. 30, Birkhäuser, Basel, 1997.

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

[24] P. Pedregal and J. Tiago, *A new existence result for autonomous non convex one-dimension optimal control problems*, J. Optim. Theory App., 134 (2007), pp. 241–255.

[25] J. P. Raymond, *Existence theorems in optimal control theory without convexity assumptions*, J. Optim. Theory Appl., 67 (1990), pp. 109–132.

[26] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[27] R. T. Rockafellar, *Existence theorems for general control problems of Bolza and Lagrange*, Adv. Math., 15 (1975), pp. 312–333.

[28] T. Roubicek, *Relaxation of optimal control problems coercive in Lp-spaces*, in Modelling and Optimization of Distributed Parameter Systems with Applications to Engineering, K. Malanowski, Z. Nahorski, and M. Peszynska, eds., Chapman & Hall, London, 1996, pp. 270–277.

[29] T. Roubicek, *Relaxation in Optimization Theory and Variational Calculus*, W. De Gruyter, Berlin, 1997.

[30] T. Roubicek and W. H. Smith, *Existence of solutions of certain nonconvex optimal control problems governed by nonlinear integral equations*, Optimization, 42 (1997), pp. 91–108.

[31] T. Roxin, *The existence of optimal controls*, Michigan Math. J., 9 (1962), pp. 109–119,

[32] J. A. Shohat and J. D. Tamarkin, *The Problem of Moments*, Math. Surveys Monogr. 1, AMS, Providence, RI, 1970.

[33] J. Warga, *Relaxed variational problems*, J. Math. Anal. Appl., 4 (1962), pp. 111–128.

[34] L. C. Young, *Generalized curves and the existence of an attained absolute minimum in the calculus of variations*, C.R. Soc. Sci. Lett. Varsovie Classe III, 30 (1937), pp. 212–234.

[35] L. C. Young, *Generalized surfaces in the calculus of variations*, Ann. of Math. (2), 43 (1942), pp. 84–103.

[36] L. C. Young, *Generalized surfaces in the calculus of variations*, II, Ann. of Math. (2), 43 (1942), pp. 530–544.

[37] L. C. Young, *Lectures on Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, 1969.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 2, pp. 438–462

© 2009 Society for Industrial and Applied Mathematics

ON A STOCHASTIC, IRREVERSIBLE INVESTMENT PROBLEM[*]

MARIA B. CHIAROLLA[†] AND ULRICH G. HAUSSMANN[‡]

Abstract. The productive sector of the economy, represented by a single firm employing labor to produce the consumption good, is studied in a stochastic continuous time model on a finite time interval. The firm must choose the optimal level of employment and capital investment in order to maximize its expected total profits. In this stochastic control problem the firm's capacity is modeled as an Itô process controlled by a monotone process, possibly singular, that represents the cumulative real investment. It is optimal to invest when the shadow value of installed capital exceeds the capital's replacement cost; this threshold is the free boundary of a related optimal stopping problem which we recast as a stopping problem without integral cost, similar to the American option problem. Then, under a regularity condition, we characterize the free boundary as the unique solution of a nonlinear integral equation.

Key words. irreversible investment, singular stochastic control, moving free boundary, optimal stopping, instantaneous stopping equation

AMS subject classifications. 91B28, 91B70, 93E20, 60G40

DOI. 10.1137/070703880

1. Introduction.

Irreversible investment problems have been studied widely in the economic literature; cf. [10] and the references therein. In these models, the producers of the goods, the firms, make decisions regarding labor levels and capital investment strategies. Most of the models are restricted to infinite horizon. For example, Abel and Eberly [1] provide an explicit solution of an irreversible investment problem in a continuous time Markovian setting, where the control processes are investment rates. Oksendal [15] considers a class of infinite horizon capital accumulation problems within a relatively large Markovian setting. Bertola [4], [5], Alvarez [2], and Riedel and Su [18] propose models with deterministic dynamics and profit rate influenced by a stochastic parameter process. [4], [5] exploit the connection with the optimal stopping problem of deciding when capital should be installed, whereas [18] uses a connection with backward stochastic differential equations while allowing both infinite and finite horizon. [2] allows a more general stochastic parameter process than [5].

In the mathematical economics literature some reversible investment problems are formulated as singular stochastic control problems. We cite, among others, Guo and Pham [11] and Merhi and Zervos [14] in the infinite horizon case and Hamadene and Jeanblanc [12] in the finite horizon case.

A more extensive review can be found in [6], where irreversible investment problems and their corresponding optimal stopping problems are linked, respectively, to

[*] Received by the editors September 27, 2007; accepted for publication (in revised form) August 21, 2008; published electronically February 13, 2009. A previous version of this paper circulated as SIIP.pdf. This work was supported by the Natural Sciences and Engineering Research Council of Canada under grant 88051 and by the Program for Cultural and Scientific Cooperation between Università di Roma "La Sapienza" and the University of British Columbia.

http://www.siam.org/journals/sicon/48-2/70388.html

[†] Corresponding author. Dipartimento di Matematica per le Dec. Econ. Finanz. e Assic., Facoltà di Economia, Università degli Studi di Roma "La Sapienza," via del Castro Laurenziano 9, 00161 Roma, Italy (maria.chiarolla@uniroma1.it).

[‡] Department of Mathematics, University of British Columbia, 1984 Mathematics Road, Vancouver, BC, V6T 1Z2 Canada (uhaus@math.ubc.ca).

the *q theory of investment* and to the theory of options, i.e., *the option value of waiting.* Of particular interest is the work of Baldursson and Karatzas [3]. They show that the solution of the myopic investor's problem, which can be solved in terms of stopping rules involving the Snell envelope, leads to a solution of the "social planner's" problem; the latter is equivalent to the optimal irreversible investment problem facing a firm. These results motivated us to study in [6], [7] an irreversible investment problem that is closely linked to the present work. Here we extend the setting of [6] in three ways; that is, we endogenize the choice of labor input, we incorporate a scrap value of the production facility by adding a terminal payoff in the control problem, and we obtain the existence result for the general case of time dependent parameters. Moreover, since the explicit expression of the free boundary obtained in [6] was incorrect and the corrections required were only outlined in [7], in this paper we provide the details which amount to finding an integral equation for the free boundary by transforming the problem into one that looks like an American option problem (cf. [13], [17]). Such a device is new and of interest in its own right.

We introduce a model of a firm that produces one good using labor $L(t)$ and utilizing production capacity (capital) $C(t)$, a geometric Brownian motion controlled by a nondecreasing process $\nu(t)$ representing the cumulative investment. We postulate a production function $R(C, L)$ and a given wage process $w(t)$. The firm chooses L and ν to maximize profit over a finite time interval, i.e.,

$$\sup_{\nu,L} E\bigg\{ \int_0^T e^{-\int_0^t \mu_F(r)\,dr}[R(C(t), L(t)) - w(t)L(t)]\,dt \\ + e^{-\int_0^T \mu_F(r)\,dr} G(C(T)) - \int_{[0,T)} e^{-\int_0^t \mu_F(r)\,dr}\,d\nu(t)\bigg\},$$

where G is the scrap value of the production facility and $\mu_F(t)$ is a random discount factor.

In section 2 we formulate the problem precisely, and using convex analysis we find a function $I^{R^A(C,\cdot)}(w)$ that will provide the optimal labor for given C and w. This reduces the problem faced by the firm to a maximization over ν only. In section 3 we establish existence and uniqueness of the optimal investment process $\hat{\nu}$. Under Markovian assumptions and some additional restrictions we show that $\hat{\nu}(t)$ is continuous on $(0, T]$ with a possible initial jump and is singular as it activates at the free boundary where the shadow value of installed capital exceeds the capital's replacement cost. The proof of continuity of $\hat{\nu}$ is new and does not require knowledge of the boundary; on the contrary in [6] it was deduced from the continuity of the incorrect boundary. In section 4 we find an algorithm to find this free boundary when the production function is of Cobb–Douglas type and the scrap value is constant. To this end we find an optimal stopping problem with no integral cost or scrap value whose solution is given by the above free boundary. The similarity with the American option problem is exploited to find an integral equation for the free boundary if the latter is continuous. We conclude with a numerical example that compares the free boundary to the incorrect one of [6]. The appendix contains a technical result on convex analysis.

2. The firm's investment problem. In this paper the firm represents the productive sector of an economy with finite horizon T modeled on a complete probability space $(\Omega, \mathcal{F}, P)$ with filtration $\{\mathcal{F}_t : t \in [0, T]\}$, which is the usual augmentation of the filtration generated by an exogenous n-dimensional Brownian motion

$\{W(t) : t \in [0,T]\}$. We work with an n-dimensional Brownian motion as we then apply the results in [8], where a two-dimensional Brownian motion is used. The firm produces a single kind of perishable consumption good at rate $R(C,L)$ when its capacity is C and it employs L units of labor.

The capital invested on the time interval $[0,t]$ for research, product development, and plant retooling or expansion is denoted by a process $\nu(t)$ a.s. finite, left-continuous with right limits, nondecreasing, and adapted. The irreversibility of investment is expressed in the nondecreasing nature of ν. The corresponding capacity is denoted by $C^{y,\nu}(t)$ and is assumed to satisfy

$$\begin{cases} dC^{y,\nu}(t) = C^{y,\nu}(t)[-\mu_C(t)dt + \sigma_C^\top(t)dW(t)] + f_C(t)d\nu(t), & t \in [0,T), \\ C^{y,\nu}(0) = y \geq 0, \end{cases} \tag{2.1}$$

where μ_C, σ_C, and f_C are given bounded, measurable, adapted processes, f_C being continuous with $0 < k_f \leq f_C \leq \kappa_f$ and $\mu_C \geq 0$. f_C is a conversion factor in the sense that each unit of new investment is converted into f_C units of capacity (it includes, for example, the cost of raising new equity).

It is convenient to define

$$C^o(t) := C^{1,0}(t) \quad \text{and} \quad \overline{\nu}(t) := \int_{[0,t)} \frac{f_C(s)}{C^o(s)}\, d\nu(s). \tag{2.2}$$

Then $C^o(t) = e^{-\int_0^t \mu_C(r)dr}\mathcal{M}_0(t)$, where, for every $s \in [0,T]$, $\mathcal{M}_s$ is the exponential martingale

$$\mathcal{M}_s(t) = e^{[\int_s^t \sigma_C^\top(r)dW(r) - \frac{1}{2}\int_s^t \|\sigma_C(r)\|^2 dr]}, \qquad t \in [s,T], \tag{2.3}$$

and $E\{[\mathcal{M}_s(t)]^p\} < \infty$ for any p. Notice that C^o represents the decay of a unit of initial capital in the absence of investment and we have $C^{y,\nu}(t) = [y + \overline{\nu}(t)]\, C^o(t)$.

Recall from convex analysis (cf. [19]) that if u is a function $\mathbb{R}^n \mapsto [-\infty, \infty)$, then the (effective) domain of u is $\mathrm{dom}(u) := \{x | u(x) > -\infty\}$ and $\mathrm{im}(u) := u(\mathrm{dom}(u))$. The function u is (strictly) concave if it is (strictly) concave on $\mathrm{dom}(u)$ (assumed to be nonempty). This makes the function a proper, concave function in the terminology of convex analysis. The supergradients of u at x are all $y \in \mathbb{R}^n$ such that for all z one has $u(z) - u(x) \leq (z-x)^\top y$. The set of all supergradients of u at x is called the supergradient set or the superdifferential and is denoted by $\partial u(x)$. The (effective) domain of $\partial u(x)$ is $\mathrm{dom}(\partial u) := \{x | \partial u(x) \neq \emptyset\}$. The (concave) conjugate function of u is defined as

$$u^*(y) := \inf_{x \in \mathbb{R}^n} \{x^\top y - u(x)\}.$$

Also, a set B is affine if it is the translate of a subspace of $\mathbb{R}^n$, including $\{0\}$ and $\mathbb{R}^n$; $\mathrm{int}(B)$ denotes the interior of the set B, $\mathrm{cl}(B)$ denotes its closure, $\mathrm{aff}(B)$ denotes its affine hull (i.e., the smallest affine set containing B), and $\mathrm{ri}(B)$ denotes the relative interior of B, i.e., the interior of B relative to $\mathrm{aff}(B)$.

The production function $R(C,L)$ of the firm is a function of the capacity C and of labor L employed. For $R(C,L)$ we make the following assumptions; cf. [9].

Assumption 1.

(i) $R : \mathbb{R}^2 \mapsto [-\infty, \infty)$ is upper semicontinuous, concave, and nondecreasing, and $[0,\infty)^2 \subset \mathrm{dom}(R)$.

(ii) R is continuous and nonnegative on $[0,\infty)^2$.

(iii) R is twice continuously differentiable on $\mathrm{int}(\mathrm{dom}(R))$.
It follows that R is a closed proper concave function such that its supergradient is $\partial R = \nabla R$ on $\mathrm{int}(\mathrm{dom}(R)) \subset \mathrm{dom}(\partial R)$.

Let $\kappa_L \in (0, \infty)$ be the labor supply and set $A := [0, \kappa_L]$ and $\tilde{A} = [0, \infty) \times A$.

Assumption 2. R is strictly increasing, strictly concave on $\mathrm{int}(\mathrm{dom}(R)) \cap \tilde{A}$.

Assumption 3. $\lim_{C\to\infty} \inf_{L\in A} R_C(C, L) = 0$.

Given a finite, constant, upper bound $\kappa_w > 0$ on possible wages, for each fixed wage rate $w \in [0, \kappa_w]$ and for each fixed C the manager of the firm will first choose $L \in A$ to maximize profits. This leads to a reduced production function $\tilde{R}(C, w)$ that can be formulated in terms of conjugate functions as follows. Let $R^A(C, \cdot)$ denote $R(C, \cdot)$ modified as $-\infty$ off A and set

$$\tilde{R}(C, w) := \sup_{L\in A}\{R(C, L) - wL\} = \sup_{L\in\mathbb{R}}\{R^A(C, L) - wL\}.$$

Assumptions 1 and 2 imply that [9, Proposition 3.2] can be applied with $u(\cdot) = R(C, \cdot)$, $A = [0, \kappa_L]$, $0^+A = \{0\}$, and $u_A(\cdot) = R^A(C, \cdot)$ to obtain the optimal L as the supergradient of the concave conjugate function of $R^A(C, \cdot)$ denoted by $I^{R^A(C,\cdot)}(w)$ in [9].

The reduced production function $\tilde{R}(C, w)$ is the negative of the concave conjugate of $R^A(C, \cdot)$; hence it is convex in w and (cf. Proposition 5.1 in the appendix) strictly concave in C. We can establish a growth condition as follows. For any $\varepsilon > 0$,

$$\sup_{C\geq 0}\{\tilde{R}(C, w) - \varepsilon C\}$$

is the negative of the concave conjugate of $R^{\tilde{A}}(\cdot, \cdot)$, i.e., of $R(\cdot, \cdot)$ extended as $-\infty$ off $\tilde{A}$. Hence it is continuous on $\mathrm{ri}(\mathrm{dom}((R^{\tilde{A}})^*))$, i.e., on $(0, \infty) \times (-\infty, \infty)$, by [9, Proposition 3.3] (note that $0^+\tilde{A} = [0, \infty) \times \{0\}$ and with this constraint set the proposition requires Assumptions 1–3). It follows that

$$\sup_{C\geq 0}\max_{w\in[0,\kappa_w]}\{\tilde{R}(C, w) - \varepsilon C\} = -\min_{w\in[0,\kappa_w]}(R^{\tilde{A}})^*(\varepsilon, w) := \kappa_\varepsilon < \infty \quad \text{for every } \varepsilon > 0, \tag{2.4}$$

where κ_ε depends on κ_L, κ_w, and ε. This is the growth condition.

Once the reduced production function $\tilde{R}(C, w)$ is obtained, given a predictable, $[0, \kappa_w]$-valued wage process $w(t, \omega)$, the manager of the firm chooses the investment $\nu(t, \omega)$ so as to maximize the expected total discounted production profits plus scrap value net of investment; i.e., he maximizes

$$\mathcal{J}_{0,y}(\nu) := E\bigg\{\int_0^T e^{-\int_0^t \mu_F(r)\,dr}\tilde{R}(C^{y,\nu}(t), w(t))\,dt + e^{-\int_0^T \mu_F(r)\,dr}G(C^{y,\nu}(T)) - \int_{[0,T)} e^{-\int_0^t \mu_F(r)\,dr}\,d\nu(t)\bigg\}. \tag{2.5}$$

Here $G(C^{y,\nu}(T))$ is a *scrap value* associated with the firm at time T and $\int_0^t \mu_F(r)\,dr$ is the manager discount factor, not to be confused with the risk neutral discount rate that usually occurs in papers including financial markets, which we do not have. Manager discount rate may be quite different, as managers are individuals and may have their own preferences, not necessarily rational.

We assume that μ_F is a bounded, nonnegative, measurable, adapted process with $\bar{\mu} := \mu_C + \mu_F \geq \varepsilon_o > 0$ a.s., and $G : [0, \infty) \mapsto [0, \infty)$ is a concave, nondecreasing, continuously differentiable function such that

$$\lim_{C \to \infty} G'(C) = 0 \quad \text{and} \quad G'(0)\, f_C(T) \leq 1 \ \text{ a.s.} \tag{2.6}$$

The firm's optimal investment problem is

$$V(0, y) := \max_{\nu \in \mathcal{S}} \mathcal{J}_{0,y}(\nu), \tag{2.7}$$

where $\mathcal{S} := \{\nu : \text{left-continuous, nondecreasing, adapted process with } \nu(0) = 0 \text{ a.s.}\}$ is the convex set of investment strategies.

Notice that the strict concavity of $\tilde{R}(\cdot, w)$, the concavity of G, and the affine nature of $C^{y,\nu}(t)$ in ν imply that $\mathcal{J}_{0,y}(\nu)$ is concave on $\mathcal{S}$; in fact, it is strictly concave since $f_C \geq k_f > 0$. Therefore, if a solution of the firm's optimal profits problem exists, then it is unique. The following estimates are needed to handle the unboundness of the reduced production function $\tilde{R}$.

PROPOSITION 2.1. *There exists a constant $K_{\mathcal{J}}$ depending on $T, \kappa_L, \kappa_w, \kappa_f, k_f$ only such that*

$$\text{(a)} \quad \mathcal{J}_{0,y}(\nu) \leq K_{\mathcal{J}}(1 + y) \quad \textit{for all} \ \ \nu \in \mathcal{S},$$

$$\text{(b)} \quad E\left\{ \int_{[0,T)} e^{-\int_0^t \mu_F(r)\,dr}\, d\nu(t) \right\} \leq 2K_{\mathcal{J}}(1 + y) \quad \textit{if} \ \ \mathcal{J}_{0,y}(\nu) \geq 0.$$

Proof. According to (2.4) and (2.6), for every $\varepsilon > 0$ there exists κ_ε, depending on ε, such that $\tilde{R}(C, w) \leq \kappa_\varepsilon + \varepsilon C$ and similarly $G(C) \leq \kappa_\varepsilon + \varepsilon C$.

We now have

$$\begin{aligned}
\mathcal{J}_{0,y}(\nu) \ &\leq \ E\bigg\{ \int_{[0,T)} e^{-\int_0^t \mu_F(r)\,dr}\Big(\kappa_\varepsilon + \varepsilon C^o(t)(y + \overline{\nu}(t))\Big)\, dt \\
&\qquad\qquad + e^{-\int_0^T \mu_F(r)\,dr}\left(\kappa_\varepsilon + \varepsilon C^o(T)\left(y + \int_{[0,T)} d\overline{\nu}(t)\right)\right) \\
&\qquad\qquad - \int_{[0,T)} e^{-\int_0^t \mu_F(r)\,dr}\, d\nu(t)\bigg\} \\
&\leq \ (\kappa_\varepsilon + \varepsilon\, y)(T + 1) + E\bigg\{ \int_{[0,T)} e^{-\int_0^t \mu_F(r)\,dr}\bigg[\varepsilon e^{\int_0^t \mu_F(r)\,dr} \frac{f_C(t)}{C^o(t)} \\
&\qquad \times E\bigg\{ \int_{[t,T)} e^{-\int_0^s \mu_F(r)\,dr} C^o(s)\, ds + e^{-\int_0^T \mu_F(r)\,dr} C^o(T) \Big| \mathcal{F}_t \bigg\} - 1\bigg] d\nu(t)\bigg\},
\end{aligned}$$

where we have interchanged the order of integration and used the fact that $C^o(t)$ is the product of a process bounded by 1 and an exponential martingale, i.e., $C^o(t) = e^{-\int_0^t \mu_C(s)ds}\mathcal{M}_0(t)$. Then the square bracket above is less than or equal to $\varepsilon(T - t + 1) f_C(t) - 1$, hence less than or equal to $-1/2$ for ε sufficiently small since $f_C(t) \leq \kappa_f$ a.s.

Now parts (a) and (b) of the proposition follow. $\square$

Remark 2.2. We point out that Proposition 2.1 still holds if, instead of (2.6), we assume that the scrap value $G : [0,\infty) \mapsto [0,\infty)$ is a concave, nondecreasing, continuously differentiable function with subaffine growth, i.e.,

$$G(C) \le a_o + b_o C, \qquad G'(0)\, f_C(T) \le 1, \qquad b_o\,\kappa_f < 1, \tag{2.8}$$

with a_o and b_o nonnegative constants.

In fact, by using $C^o(s)/C^o(t) = e^{-\int_t^s \mu_C(r)dr}\mathcal{M}_t(s)$ for $t \le s \le T$ and the third part of (2.8), we obtain

$$\mathcal{J}_{0,y}(\nu) \le (\kappa_\varepsilon + \varepsilon\, y)T + a_o + b_o\, y + E\left\{ \int_{[0,T)} e^{-\int_0^t \mu_F(r)\,dr}[(\varepsilon T + b_o)f_C(t) - 1]\, d\nu(t) \right\}$$

with $[(\varepsilon T + b_o)f_C(t) - 1] < -(1 - b_o\kappa_f)/2$ if $\varepsilon < (1 - b_o\kappa_f)/(2T\kappa_f)$. Then $\mathcal{J}_{0,y}(\nu) \le K'_{\mathcal{J}}(1 + y)$ with $K'_{\mathcal{J}}$ depending only on $a_o, b_o, T, \kappa_L, \kappa_w$ and the bounds on f_C. Similarly, $E\int_{[0,T)} e^{-\mu_F(t)}\, d\nu(t) \le 2K'_{\mathcal{J}}\frac{(1+y)}{(1-b_o\kappa_f)}$ if $\mathcal{J}_{0,y}(\nu) \ge 0$. The result follows with $K_{\mathcal{J}} := \frac{K'_{\mathcal{J}}}{(1-b_o\kappa_f)}$.

3. Solution of the optimal investment problem. We now aim to generalize the optimal investment existence theorem of [6]. Notice that in [6] we had constant parameters and no labor. By using some results in [3], we shall find the solution of (2.7) via the optimal stopping problem naturally associated with it. In fact, we define the opportunity cost of not investing until time t, when the initial capacity is y, by

$$\begin{aligned}\zeta^{y,T}(t) := \int_0^t e^{-\int_0^s \mu_F(r)\,dr} C^o(s)\tilde{R}_C(yC^o(s), w(s))\, ds + e^{-\int_0^t \mu_F(r)\,dr}\frac{C^o(t)}{f_C(t)}\mathbf{1}_{\{t<T\}} \\ + e^{-\int_0^T \mu_F(r)\,dr} C^o(T)G'(yC^o(T))\mathbf{1}_{\{t=T\}};\end{aligned}$$

hence we define the optimal stopping problem

$$Z^{y,T}(t) := \operatorname*{ess\,inf}_{\tau\in\Upsilon[t,T]} E\{\zeta^{y,T}(\tau)|\mathcal{F}_t\}, \tag{3.1}$$

where $\Upsilon[t,T]$ is the set of stopping times τ with values in $[t,T]$. We denote by $\mathcal{Z}^{y,T}(\cdot\,)$ a modification of $Z^{y,T}(\cdot\,)$ with right-continuous with left limits paths; then we set

$$\tau^*(0,y) := \inf\{s \in [0,T) : \mathcal{Z}^{y,T}(s) = \zeta^{y,T}(s)\} \wedge T, \tag{3.2}$$

$$\overline{\nu}^y(t) := [\sup\{z \ge y : \tau^*(0,z+) < t\} - y]^+ \text{ if } t > 0, \text{ and } \overline{\nu}^y(0) = 0. \tag{3.3}$$

Thus $\overline{\nu}^y$ (modulo a shift) is $\tau^*(0,\cdot)$'s left-continuous inverse. Notice that $\tau^*(0,y)$ is nondecreasing in y a.s.; cf. [3, Lemma 1]. Also, if $\hat{y}(0) := \sup\{z \ge 0 : \tau^*(0,z) = 0\}$, then $\overline{\nu}^y(0+) = \max\{\hat{y}(0) - y, 0\} := [\hat{y}(0) - y]^+$; i.e., expanding capacity up to level $\hat{y}(0)$ is the optimal strategy at time 0+.

Then we have the following theorem.

THEOREM 3.1. *Assume that w is a continuous process with values in $[0,\kappa_w]$, and either condition* (2.6) *or condition* (2.8) *holds. For y fixed, set*

$$\hat{\nu}(t) := \int_{[0,t)} \frac{C^o(s)}{f_C(s)}\, d\overline{\nu}^y(s). \tag{3.4}$$

Then the following hold:

(i) *$\hat\nu$ is the unique optimal control of the investment problem* $\max_{\nu\in\mathcal{S}}\mathcal{J}_{0,y}(\nu)$.

(ii) $E\|\hat\nu\|_T \le 2K_{\mathcal{J}}\,(1+y)\max_{t,\omega} e^{\int_0^t \mu_F(r)\,dr}$.

(iii) *If $C^{y,\hat\nu}(T)\equiv 0$, then $y=0$ and $\hat\nu\equiv 0$; moreover, a.e. a.s.,*

$$e^{\int_0^t \mu_F(r)\,dr} f_C(t) E\left\{\int_t^T e^{-\int_0^s \mu_F(r)\,dr}\tilde R_C(yC^o(s),w(s))\frac{C^o(s)}{C^o(t)}\,ds \right. \\ \left. + e^{-\int_0^T \mu_F(r)\,dr} G'(yC^o(T))\frac{C^o(T)}{C^o(t)}\right|\mathcal{F}_t\Bigg\} \le 1. \tag{3.5}$$

Proof. Since $0\le \mathcal{J}_{0,y}(0)$, (ii) follows easily from (b) of Proposition 2.1 and the nondecreasing paths of $\hat\nu$.

Part (iii) follows from $C^{y,\hat\nu}(T) = [y+\overline{\nu}^y(T)]\,C^o(T)$. To show (3.5) we calculate the Gâteaux derivative of $\mathcal{J}_{0,y}(\cdot)$ at 0 in the direction ν:

$$d\mathcal{J}_{0,y}(0;\nu) = E\int_{[0,T)}\left[f_C(t)\,E\left\{\int_t^T e^{-\int_0^s \mu_F(r)\,dr}\tilde R_C(yC^o(s),w(s))\frac{C^o(s)}{C^o(t)}\,ds\right.\right. \\ \left.\left. + e^{-\int_0^T \mu_F(r)\,dr} G'(yC^o(T))\frac{C^o(T)}{C^o(t)}\right|\mathcal{F}_t\Bigg\} - e^{-\int_0^t \mu_F(r)\,dr}\right] d\nu(t).$$

If 0 is optimal, then $d\mathcal{J}_{0,y}(0;\nu)\le 0$; i.e., (3.5) holds.

For (i) we transform the problem into "the social planning problem" of [3] and verify that the assumptions are met. This is done in [6] for a simpler case. In [3] the following variables occur: Π, γ_o, G_o (denoted γ, G in [3]), p, g, which we now define as

$$\begin{cases} \Pi(t,y) &:= & e^{-\int_0^t \mu_F(r)\,dr}[\tilde R(yC^o(t),w(t)) - \tilde R(C^o(t),w(t))], \\ \gamma_o(t) &:= & e^{-\int_0^t \mu_F(r)\,dr}C^o(t)f_C(t)^{-1}, \\ G_o(y) &:= & e^{-\int_0^T \mu_F(r)\,dr}[G(yC^o(T)) - G(C^o(T))], \\ p(t,y) &:= & e^{-\int_0^t \mu_F(r)\,dr}C^o(t)\tilde R_C(yC^o(t),w(t)), \\ g(y) &:= & e^{-\int_0^T \mu_F(r)\,dr}C^o(T)G'(yC^o(T)). \end{cases}$$

Notice that Proposition 5.1 in the appendix implies that

$$\tilde R_C(yC^o(t),w(t)) = R_C(yC^o(t), I^{R^A(yC^o(t),\cdot)}(w(t))) \tag{3.6}$$

with $I^{R^A(C,\cdot)}(w)$ continuous in (C,w). Hence $p(t,y)$ is continuous since w is. Moreover, p is strictly decreasing in y since $\tilde R$ is strictly concave in its first variable.

Now J of [3] becomes

$$J(y,\overline{\nu}) := E\int_0^T \Pi(t,y+\overline{\nu}(t))\,dt + G_o(y+\overline{\nu}(T)) - \int_{[0,T)}\gamma_o(t)\,d\overline{\nu}(t),$$

with $\overline{\nu}$ defined in (2.2). Then $\mathcal{J}_{0,y}(\nu) = J(y,\overline{\nu}) + \mathcal{J}_{0,1}(0)$. Hence if we find $\overline{\nu}$ which maximizes $J(y,\cdot\,)$, then the corresponding ν will be optimal for $\mathcal{J}_{0,y}$. It remains to show that $p\ (=\Pi_y)$, $g\ (=G_o')$, and γ_o satisfy (2.1), (2.3), (2.4), (3.2), (3.3), and (4.4) of [3]. In fact, (3.3) of [3] follows from the second half of our (2.6), and the integrability of $p(\cdot\,,y)$ is required only for $y>0$. We do not have (4.4) of [3] since $\tilde R(C,w)$ and $G(C)$ are not bounded, but $\mathcal{J}_{0,y}(\nu)\le K_{\mathcal{J}}(1+y)$ by Proposition 2.1 (Remark 2.2, respectively), and this is all that is needed in [3, definitions (4.3), (4.3)′]. So the arguments of [6] apply to establish the result. □

Remark 3.2. If R is of Cobb–Douglas type with zero shift, $R(C,L) = \frac{1}{\alpha\beta} C^\alpha L^\beta$ with $\alpha, \beta > 0$, and $\alpha + \beta < 1$, then $I^{R^A(C,\cdot)}(w) = (\frac{C^\alpha}{\alpha w})^{\frac{1}{1-\beta}}$ for C near zero, and $R_C(C, I^{R^A(C,\cdot)}(w)) = C^{\frac{\alpha+\beta-1}{1-\beta}}(\beta(\alpha w)^{\frac{\beta}{1-\beta}})^{-1} \to \infty$ as $C \to 0$, for any $w \in [0, \kappa_w]$. Since $\tilde{R}_C(yC^o(s), w(s)) = R_C(yC^o(s), I^{R^A(yC^o(s),\cdot)}(w(s)))$ (cf. (3.6)), $\lim_{y\downarrow 0} d\mathcal{J}_{0,y}(0;\nu) = \infty$ by monotone convergence (cf. the proof of (3.5)), so (3.5) fails; in fact, $C^{y,\hat{\nu}}(t) > 0$ for all $t > 0$.

On the other hand, if μ_F and μ_C are constant and if $R_C(0,L) = R_C(0)$ is labor independent (e.g., if (3.8)(i) below holds), then (3.5) reduces to

$$\text{(3.7)}\quad \tilde{R}_C(0)\left(\frac{1 - e^{-(\mu_F+\mu_C)(T-t)}}{\mu_F + \mu_C}\right) + G'(0)e^{-(\mu_F+\mu_C)(T-t)} \le \frac{1}{f_C(t)} \qquad \text{a.e. a.s.}$$

So if $y = 0$, then $\hat{\nu} \equiv 0$; i.e., the production facility will not be built if the marginal return at zero capacity (the left-hand side of (3.7)), hence at any capacity greater than zero, is never greater than the marginal cost of new capacity (the right-hand side).

We now make some "Markovian" restrictions in order to obtain the continuity of $\hat{\nu}(t)$.

Assumption-[M].

$$\text{(3.8)}\qquad \begin{aligned} &\text{(i)} \quad R(C,L) = R^1(C) + R^2(L), \\ &\text{(ii)} \quad \mu_F, \mu_C, \sigma_C, f_C \quad \text{are constant.} \end{aligned}$$

Such an additive production function make sense if, for example, C represents technology, then buying or finding new technology may increase productivity with little change in labor. Observe that now $\tilde{R}_C(yC^o(t), w(t)) = R^1_C(yC^o(t))$ is continuous irrespective of w. Also recall that Assumptions 2 and 3 hold for R^1 and they do not assume $R^1_C(0) = \infty$, as does the Inada condition.

Assumption-[G].

$$\text{(3.9)}\qquad \begin{aligned} &\text{(i)} \quad G \in C^3([0,\infty)), \\ &\text{(ii)} \quad |G''(C)| \le \kappa_G(1 + |C|^{k_G}) \quad \text{for some constant } \kappa_G \text{ and some (possibly} \\ &\qquad \text{negative) constant exponent } k_G, \\ &\text{(iii)} \quad R^1_C(C) - (\mu_C + \mu_F)G'(C) \\ &\qquad\qquad + (\|\sigma_C\|^2 - \mu_C)CG''(C) + \|\sigma_C\|^2 C^2 G'''(C)/2 \ge 0. \end{aligned}$$

Under Assumption-[M] we are in the setting of section 4 of [6] but with scrap value. Then the capacity process starting at time t from $y > 0$ and controlled by ν,

$$\begin{cases} dC^{t,y,\nu}(s) = C^{t,y,\nu}(s)[-\mu_C ds + \sigma_C^\top dW(s-t)] + f_C\, d\nu(s-t), & s \in (t,T], \\ C^{t,y,\nu}(t) = y, \end{cases}$$

is $\mathcal{F}_{s-t}$-adapted and time-homogeneous since it may be identified with $C^{y,\nu}(s-t)$, the process starting at time zero from y. It follows that the corresponding profit may be written in terms of $C^{y;\nu}$, that is,

$$\mathcal{J}_{t,y}(\nu) = E\biggl\{ \int_0^{T-t} e^{-\mu_F\, s}\tilde{R}(C^{y,\nu}(s))ds + e^{-\mu_F\,(T-t)}G(C^{y,\nu}(T-t))$$

$$- \int_{[0,T-t)} e^{-\mu_F\, s}\, d\nu(s) \Big\}.$$

Hence the associated optimal stopping problem is $Z^{y,T-t}(0)$ (cf. (3.1)).

We define $v(t,y) := \mathcal{Z}^{y,T-t}(0)$, a right-continuous with left limits modification of $Z^{y,T-t}(0)$; hence up to a null set

$$\begin{aligned} v(t,y) = \inf_{\tau \in \Upsilon[0,T-t]} E\Big\{ \int_0^\tau e^{-\mu_F\, s} C^o(s) R^1_C(y\, C^o(s))\, ds + e^{-\mu_F\, \tau} C^o(\tau) \frac{1}{f_C} \mathbf{1}_{\{\tau < T-t\}} \\ + e^{-\mu_F\, (T-t)} C^o(T-t) G'(y\, C^o(T-t)) \mathbf{1}_{\{\tau = T-t\}} \Big\}. \end{aligned} \tag{3.10}$$

Now a generalization of (4.8) of [6] to the case with scrap value implies that

$$v(t,y) = \frac{\partial}{\partial y} V(t,y); \tag{3.11}$$

i.e., v is the shadow value of installed capital, and

$$\tau^*(t,y) := \inf\{ s \in [0, T-t) : \mathcal{Z}^{y,T-t}(s) = \zeta^{y,T-t}(s) \} \wedge (T-t) \tag{3.12}$$

is optimal for $v(t,y)$ (cf. the second half of (2.6), the second part of (2.8), (3.2), and [3]).

Notice that if $\tau^*(t,y) = 0$ for some (t,y) with $t < T$, then $\tau^*(t,z) = 0$ for all $z < y$ since R^1_C and G' are nonincreasing. Hence we may define on $\{t < T\}$ the "boundary"

$$\hat{y}(t) := \sup\{ z \geq 0 : \tau^*(t,z) = 0 \}. \tag{3.13}$$

The following result is new and was not contained in [6], where the continuity of $\hat{\nu}$ was an obvious consequence of the explicit expression of the boundary, which unfortunately was incorrect.

Proposition 3.3. *Under Assumption-*[M] *and Assumption-*[G],

- v *is* $\begin{cases} \text{nonincreasing in } y \text{ for each } t \\ \text{nonincreasing in } t \text{ for each } y \end{cases}$ *and* $v \leq \frac{1}{f_C}$;
- $\hat{y}(t)$ *is nonincreasing;*
- *the optimal investment process* $\hat{\nu}(t)$ *is continuous on* $0 < t \leq T$, *except perhaps for an initial jump.*

Proof. We set $Y^y(t) := y\, C^o(t)$ under a new measure $Q \sim P$ with $\frac{dQ}{dP} = \exp\left\{ \sigma_C^\top W(T) - \frac{1}{2} \|\sigma_C\|^2 T \right\}$. Then $W(t) - \sigma_C t$ is a Wiener process under Q and $B^Q(t) := \frac{1}{\|\sigma_C\|} \sigma_C^\top (W(t) - \sigma_C t)$ is a one-dimensional Brownian motion in terms of which we have

$$\begin{cases} dY^y(t) = Y^y(t) \left[(\|\sigma_C\|^2 - \mu_C)\, dt + \|\sigma_C\|\, dB^Q(t) \right], & t \in (0,T], \\ Y^y(0) = y. \end{cases} \tag{3.14}$$

With $\bar{\mu} := \mu_C + \mu_F$, we may write

$$\begin{aligned} v(t,y) = \inf_{\tau \in \Upsilon[0,T-t]} E^Q\Big\{ \int_0^\tau e^{-\bar{\mu}\, s} R^1_C(Y^y(s))\, ds + e^{-\bar{\mu}\, \tau} \frac{1}{f_C} \mathbf{1}_{\{\tau < T-t\}} \\ + e^{-\bar{\mu}\, \tau} G'(Y^y(\tau)) \mathbf{1}_{\{\tau = T-t\}} \Big\}; \end{aligned} \tag{3.15}$$

now the arguments of Theorem 4.1 of [6] apply and show that the optimal stopping time of (3.15) may be characterized as the first time $(s, Y^y(s))$ hits the boundary of $v(t+\cdot\,,\cdot\,) = 1/f_C$, i.e.,

$$\tau^*(t,y) = \inf\left\{ s \in [0, T-t) : v(t+s, Y^y(s)) = \frac{1}{f_C} \right\} \wedge (T-t). \tag{3.16}$$

It is easy to see that v is nonincreasing in y (in fact, $v(t,y)$ is strictly decreasing in y on $y > \hat{y}(t)$) and $v \le \frac{1}{f_C}$; hence

$$\text{for each } t, \text{ the set } \quad \left\{ y > 0 : v(t,y) < \frac{1}{f_C} \right\} \quad \text{is connected.} \tag{3.17}$$

We shall now show that v is also nonincreasing in t. With $t_1 < t_2$ and $\tau \in \Upsilon[0, T-t_1]$, let $\tilde{\tau} : \tau \wedge (T-t_2)$, $\tau' := (\tau - T + t_2) \vee 0$, so $\tau = \tilde{\tau} + \tau'$. Note that $\tilde{\tau} \in \Upsilon[0, T-t_2]$ but τ' is only an $\{\mathcal{F}_{\cdot\, + T - t_2}\}$-stopping time. Let $\tilde{y} := Y^y(T-t_2)$. Then

$$\begin{aligned} v(t_1, y) = \inf_{\tau \in \Upsilon[0,T-t_1]} E^Q \bigg\{ & \int_0^{\tilde{\tau}} e^{-\bar{\mu}\, s} R_C^1(Y^y(s))\, ds + \int_{\tilde{\tau}}^{\tau} e^{-\bar{\mu}\, s} R_C^1(Y^y(s))\, ds \\ & + e^{-\bar{\mu}\, \tau} \frac{1}{f_C} \mathbf{1}_{\{\tau < T - t_2\}} + e^{-\bar{\mu}\, \tau} \frac{1}{f_C} \mathbf{1}_{\{T-t_2 \le \tau < T - t_1\}} \\ & + e^{-\bar{\mu}\,(T-t_1)} G'(Y^y(T-t_1)) \mathbf{1}_{\{\tau = T - t_1\}} \bigg\} \\ \ge \inf_{\tilde{\tau} \in \Upsilon[0,T-t_2]} E^Q \bigg\{ & \int_0^{\tilde{\tau}} e^{-\bar{\mu}\, s} R_C^1(Y^y(s))\, ds + e^{-\bar{\mu}\, \tilde{\tau}} \frac{1}{f_C} \mathbf{1}_{\{\tilde{\tau} < T - t_2\}} \\ & + \mathbf{1}_{\{\tilde{\tau} = T - t_2\}} e^{-\bar{\mu}\,(T-t_2)} \operatorname*{ess\,inf}_{0 \le \tau' \le t_2 - t_1} \bar{g}(\tilde{y}, \tau') \bigg\}, \end{aligned}$$

where

$$\begin{aligned} \bar{g}(y, \tau') := E^Q \bigg\{ & \int_0^{\tau'} e^{-\bar{\mu}\, s} R_C^1(Y^y(s))\, ds + e^{-\bar{\mu}\, \tau'} \frac{1}{f_C} \mathbf{1}_{\{\tau' < t_2 - t_1\}} \\ & + e^{-\bar{\mu}\,(t_2-t_1)} G'(Y^y(t_2 - t_1)) \mathbf{1}_{\{\tau' = t_2 - t_1\}} \bigg\} \\ \ge E^Q \bigg\{ & \int_0^{\tau'} e^{-\bar{\mu}\, s} R_C^1(Y^y(s))\, ds + e^{-\bar{\mu}\, \tau'} G'(Y^y(\tau')) \bigg\} \end{aligned}$$

since $1/f_C \ge G'(0) \ge G'(Y^y(\tau'))$. If we apply Itô's lemma to the term inside the last expectation, then Assumption-[G] implies that it dominates $G'(Y^y(0)) + \int_0^{\tau'} e^{-\bar{\mu}\, t} G''(Y^y(t)) Y^y(t) \sigma_C^\top \, dB^Q(t)$. The bound (3.9)(ii) on G'' guarantees that the stochastic integral has mean 0, so $\bar{g}(y, \tau') \ge G'(y)$. Hence

$$\begin{aligned} v(t_1, y) \ge \inf_{\tilde{\tau} \in \Upsilon[0,T-t_2]} E^Q \bigg\{ & \int_0^{\tilde{\tau}} e^{-\bar{\mu}\, s} R_C^1(Y^y(s))\, ds + e^{-\bar{\mu}\, \tilde{\tau}} \frac{1}{f_C} \mathbf{1}_{\{\tilde{\tau} < T - t_2\}} \\ & + e^{-\bar{\mu}\, \tilde{\tau}} G'(Y^y(\tilde{\tau})) \mathbf{1}_{\{\tilde{\tau} = T - t_2\}} \bigg\} = v(t_2, y); \end{aligned}$$

i.e., $t \mapsto v(t,y)$ is nonincreasing. As a consequence,

$$\text{for each } y > 0, \text{ the set } \left\{ t \geq 0 : v(t,y) < \frac{1}{f_C} \right\} \text{ is connected.} \tag{3.18}$$

Then (3.16) and $v(t,y) \leq \frac{1}{f_C}$ imply that $\tau^*(0,y)$ is the first exit time of $(s, Y^y(s))$ from the Borel set

$$\Delta := \left\{ (t,y) \in [0,T) \times (0,\infty) : v(t,y) < \frac{1}{f_C} \right\}, \tag{3.19}$$

the continuation region of problem (3.15) (see also [6, definition (4.12)]). The monotonicity properties of v established above imply that *the left boundary of* Δ *is a nonincreasing curve in the* (t,y)*-plane* and Δ *lies to the right and above this curve*; in fact, the curve is graph$(\hat{y})$.

As pointed out below (3.3), $\tau^*(0,y)$ is nondecreasing in y a.s.; we now show that, in fact, $y \mapsto \tau^*(0,y)$ *is strictly increasing a.s. on* $y > \hat{y}(0)$.

Take $x > y > \hat{y}(0)$ and $\tau^*(0,y) > s \geq 0$. Then $Y^x(s) - Y^y(s) = (x-y)C^o(s) > 0$, so that $(s, Y^y(s))$ always lies strictly below $(s, Y^x(s))$. Now the nonincreasing nature of the boundary implies that at $s = \tau^*(0,y)$, $(s, Y^y(s))$ lies in the boundary but $(s, Y^x(s))$ still lies in the interior of Δ, so $\tau^*(0,y) < \tau^*(0,x)$.

It follows that $\tau^*(0,\cdot)$'s left-continuous inverse (i.e., $\overline{\nu}^y$ modulo a shift) is continuous except possibly for an initial jump and so $\hat{\nu}$ is continuous except possibly for an initial jump. □

4. An algorithm for the free boundary. In this section we work under Assumption-[M]; we take the production function $R^1(C)$ to be of Cobb–Douglas type, i.e., $R^1(C) = \frac{1}{\alpha}C^\alpha$ with $0 < \alpha < 1$. Finally, we specify the scrap value $G(y)$ to be constant, i.e., $G(y) = a_o \geq 0$.

In order to find an algorithm for the free boundary $\hat{y}(t)$ defined in (3.13), we reformulate the optimal stopping problem (3.15) into a stopping problem with no integral cost or scrap value, as is the case for the American option problem. This is accomplished as follows.

Recall that $Y^y(t) = y\, e^{(\frac{1}{2}\|\sigma_C\|^2 - \mu_C)t + \|\sigma_C\|\, B^Q(t)}$ (see (3.14)), so we have the following useful equality:

$$e^{-\bar{\mu}r}\Big(Y^y(r)\Big)^{\alpha-1} = y^{\alpha-1}\, e^{P(\alpha-1)\, r}\, m(r), \qquad y > 0, \tag{4.1}$$

involving the Q-martingale $m(r) := e^{[-\frac{1}{2}(\alpha-1)^2\|\sigma_C\|^2\, r + (\alpha-1)\|\sigma_C\|\, B^Q(r)]}$ and the polynomial P with $P(\alpha-1) := \frac{1}{2}\|\sigma_C\|^2\,(\alpha-1)^2 + (\frac{1}{2}\|\sigma_C\|^2 - \mu_C)\,(\alpha-1) - \bar{\mu} < 0$ (cf. [6, definition (5.5)]).

Let L denote the differential generator of Y^y, i.e., $L := \frac{1}{2}\|\sigma_C\|^2\, y^2\, \partial_{yy} + (\|\sigma_C\|^2 - \mu_C) y\, \partial_y$. Let φ be the solution of $\varphi_t(t,y) + (L - \bar{\mu})\varphi(t,y) = y^{\alpha-1}$ with $\varphi(0,y) = 0$. Then φ is the $C^\infty((0,T) \times (0,\infty))$-function, strictly increasing in t, strictly decreasing and convex in y, given by

$$\varphi(t,y) = \frac{[e^{-P(\alpha-1)t} - 1]}{-P(\alpha-1)}\, y^{\alpha-1}. \tag{4.2}$$

Define

$$H(t,y) := \varphi(t,y) + \frac{1}{f_C}\, \mathbf{1}_{\{t<T\}}, \tag{4.3}$$

and introduce the optimal stopping problem

$$\hat{v}(t,y) := \inf_{\tau \in \Upsilon[0,T-t]} E^Q\Big\{e^{-\bar{\mu}\tau} H(t+\tau, Y^y(\tau))\Big\}, \qquad y > 0. \tag{4.4}$$

The definition of H and the equation for φ link $\hat{v}$ to the value function v of the optimal stopping problem (3.15); in fact, for t fixed,

$$\begin{aligned} &d[e^{-\bar{\mu}s}\varphi(t+s, Y^y(s))] \\ &= e^{-\bar{\mu}s}\Big\{[\varphi_t + (L-\bar{\mu})\varphi](t+s, Y^y(s))ds + \varphi_y(t+s, Y^y(s))Y^y(s)\|\sigma_C\|\, dB^Q(s)\Big\} \\ &= e^{-\bar{\mu}s}\Big\{(Y^y(s))^{\alpha-1}\, ds + \varphi_y(t+s, Y^y(s))Y^y(s)\|\sigma_C\|\, dB^Q(s)\Big\} \end{aligned}$$

and $E^Q\{[\mathcal{M}_0(t)]^{2\alpha-2}\} < \infty$ (cf. (2.3)) implies, for every stopping time $\tau \in \Upsilon[0, T-t]$, that

$$\begin{aligned} \varphi(t,y) &= E^Q\Big\{e^{-\bar{\mu}\,\tau}\varphi(t+\tau, Y^y(\tau)) - \int_0^\tau e^{-\bar{\mu}\,s}(Y^y(s))^{\alpha-1}\, ds\Big\} \\ &= E^Q\Big\{e^{-\bar{\mu}\,\tau}\Big[H(t+\tau, Y^y(\tau)) - \frac{1}{f_C}\mathbf{1}_{\{t+\tau<T\}}\Big] - \int_0^\tau e^{-\bar{\mu}s}(Y^y(s))^{\alpha-1} ds\Big\}; \end{aligned}$$

that is,

$$\varphi(t,y) = E^Q\Big\{e^{-\bar{\mu}\,\tau}\Big[H(t+\tau, Y^y(\tau)) - \frac{1}{f_C}\mathbf{1}_{\{t+\tau<T\}}\Big] - \int_0^\tau e^{-\bar{\mu}s}(Y^y(s))^{\alpha-1} ds\Big\}.$$

Hence, by taking the inf over $\tau \in \Upsilon[0, T-t]$, we obtain

$$\hat{v}(t,y) = v(t,y) + \varphi(t,y). \tag{4.5}$$

As a consequence, the continuation region Δ (cf. (3.19)) may be written in terms of $\hat{v}$, i.e.,

$$\Delta = \{(t,y) \in [0,T) \times (0,\infty) : \hat{v}(t,y) < H(t,y)\}, \tag{4.6}$$

and similarly for its t-section, i.e., $\Delta_t = \{y \in (0,\infty) : \hat{v}(t,y) < H(t,y)\}$.

Moreover, the optimal stopping time $\tau^*(t,y)$ of $v(t,y)$ (cf. (3.16)) is also optimal for $\hat{v}(t,y)$ since

$$\begin{aligned} \tau^*(t,y) &= \inf\Big\{s \in [0, T-t) : v(t+s, Y^y(s)) = \frac{1}{f_C}\Big\} \wedge (T-t) \\ &= \inf\{s \in [0, T-t] : \hat{v}(t+s, Y^y(s)) = H(t+s, Y^y(s))\}, \end{aligned} \tag{4.7}$$

and it is nondecreasing in y (cf. the discussion following (3.3)).

Proposition 4.1. *The value function $\hat{v}(t,y)$ of the new optimal stopping problem* (4.4) *has the following properties:*

[i]$_{\hat{v}}$ *$\hat{v}(t,\cdot)$ is nonincreasing in y for all $t \in [0,T]$;*

[ii]$_{\hat{v}}$ *$z \to \hat{v}(t, z^{\alpha-1})$ is concave on $(0,\infty)$ for all $t \in [0,T]$;*

[iii]$_{\hat{v}}$ *$\hat{v}(t,\cdot)$ is continuous in y on $(0,\infty)$ for all $t \in [0,T]$, and*

$$\hat{v}(t,x) - \hat{v}(t,y) \le \frac{e^{-P(\alpha-1)T} - 1}{-P(\alpha-1)}|x^{\alpha-1} - y^{\alpha-1}|;$$

[iv]$_{\hat{v}}$ *$(\hat{v}-H)(t,\cdot)$ is nonincreasing on $(0,\infty)$ for all $t\in[0,T)$;*

[v]$_{\hat{v}}$ *for $t\in[0,T)$, at points $y\in(0,\infty)$ where $\hat{v}_y(t,y)$ exists, one has*

$$|\hat{v}_y(t,y)| \le \frac{e^{-P(\alpha-1)t}}{-P(\alpha-1)}\,|\alpha-1|\frac{y^{\alpha-1}}{y};$$

[vi]$_{\hat{v}}$ *$\hat{v}(\cdot,y)$ is nondecreasing in t on $[0,T]$ for all $y\in(0,\infty)$ (despite the discontinuity of H at $t=T$);*

[vii]$_{\hat{v}}$ *$\hat{v}(\cdot,y)$ is continuous in t on $[0,T]$ for all $y\in(0,\infty)$;*

[viii]$_{\hat{v}}$ *$\hat{v}$ is continuous in $(t,y)\in[0,T]\times(0,\infty)$;*

[ix]$_{\hat{v}}$ *$(\hat{v}-H)(\cdot,y)$ is nonincreasing on $[0,T)$ for all $y\in(0,\infty)$.*

Proof. Property [i]$_{\hat{v}}$ follows from the analogous property of $H(t,\cdot)$.

To show [ii]$_{\hat{v}}$, recall that $H(t,y)$ is convex in y for each t fixed since $H(t,y) := \varphi(t,y) + (\frac{1}{f_C}-b_o)\,\mathbf{1}_{\{t<T\}} = \frac{[e^{-P(\alpha-1)t}-1]}{-P(\alpha-1)}\,y^{\alpha-1} + b_o + (\frac{1}{f_C}-b_o)\,\mathbf{1}_{\{t<T\}}$. It follows that $H(t,z^{\alpha-1})$ is concave in z since $0<(\alpha-1)^2<1$. Hence $z\to\hat{v}(t,z^{\alpha-1})$ is concave, being the inf of concave functions.

Now the continuity of $\hat{v}$ in y follows from the concavity of $\hat{v}(t,z^{\alpha-1})$ and the continuity of $y^{\frac{1}{\alpha-1}}$ on $(0,\infty)$. Also, for $x,y\in(0,\infty)$, we have the following estimate:

$$\begin{aligned}
(4.8)\ \hat{v}(t,x)-\hat{v}(t,y)& \\
&\le E^Q\Big\{e^{-\bar{\mu}\tau^*(t,y)}\Big[H(t+\tau^*(t,y),Y^x(\tau^*(t,y))) - H(t+\tau^*(t,y),Y^y(\tau^*(t,y)))\Big]\Big\}\\
&= E^Q\Big\{e^{-\bar{\mu}\tau^*(t,y)}\Big[\varphi(t+\tau^*(t,y),Y^x(\tau^*(t,y))) - \varphi(t+\tau^*(t,y),Y^y(\tau^*(t,y)))\Big]\Big\}\\
&= E^Q\left\{\left[\frac{e^{-P(\alpha-1)(t+\tau^*(t,y))}-1}{-P(\alpha-1)}\right]e^{P(\alpha-1)\tau^*(t,y)}\,m(\tau^*(t,y))\right\}(x^{\alpha-1}-y^{\alpha-1}),
\end{aligned}$$

where we have used the definition of P and the equality (4.1). Thus [iii]$_{\hat{v}}$ follows.

From the above computation we also obtain

$$\hat{v}(t,x)-\hat{v}(t,y) \le E^Q\left\{\frac{e^{-P(\alpha-1)t}m(\tau^*(t,y)) - e^{P(\alpha-1)\tau^*(t,y)}m(\tau^*(t,y))}{-P(\alpha-1)}\right\}(x^{\alpha-1}-y^{\alpha-1}),$$

and therefore

$$\hat{v}(t,x)-\hat{v}(t,y) \le \frac{e^{-P(\alpha-1)t}}{-P(\alpha-1)}(x^{\alpha-1}-y^{\alpha-1}), \quad \text{for } x<y,$$

since $x^{\alpha-1}-y^{\alpha-1}>0$ and $\frac{-e^{P(\alpha-1)\tau^*(t,y)}\,m(\tau^*(t,y))}{-P(\alpha-1)}<0$; whereas

$$\hat{v}(t,x)-\hat{v}(t,y) \le \frac{[e^{-P(\alpha-1)t}-1]}{-P(\alpha-1)}\,(x^{\alpha-1}-y^{\alpha-1}) = H(t,x)-H(t,y), \quad \text{for } x>y, \tag{4.9}$$

since $x^{\alpha-1}-y^{\alpha-1}<0$ and $e^{P(\alpha-1)\,r}\,m(r)$ is a Q-supermartingale. Inequality (4.9) implies [iv]$_{\hat{v}}$ and also

$$\frac{e^{-P(\alpha-1)t}}{-P(\alpha-1)}\,(\alpha-1)\frac{y^{\alpha-1}}{y} \le \hat{v}_y(t,y-), \qquad \hat{v}_y(t,y+) \le \frac{[e^{-P(\alpha-1)t}-1]}{-P(\alpha-1)}\,(\alpha-1)\frac{y^{\alpha-1}}{y};$$

therefore [v]$_{\hat{v}}$ holds.

Although $[\mathrm{vi}]_{\hat{v}}$ follows from Proposition 3.3, we give here a direct proof. Let $t_1, t_2 \in [0,T)$ with $t_1 < t_2$, take $y > 0$, and define $\tau_1 := t_2 - t_1 + \tau^*(t_2, y)$. Then $0 \le \tau_1 \le T - t_1$ is admissible at (t_1, y) and

$$\begin{aligned}
&\hat{v}(t_1,y) - \hat{v}(t_2,y)\\
&\quad \le E^Q\Big\{e^{-\bar{\mu}\tau_1}\varphi(t_1+\tau_1, Y^y(\tau_1)) - e^{-\bar{\mu}\tau^*(t_2,y)}\varphi(t_2+\tau^*(t_2,y), Y^y(\tau^*(t_2,y)))\Big\}\\
&\quad \le y^{\alpha-1}E^Q\bigg\{\frac{e^{-P(\alpha-1)(t_2+\tau^*(t_2,y))}-1}{-P(\alpha-1)}\\
&\qquad\qquad \times\Big[e^{P(\alpha-1)\,\tau_1}m(\tau_1) - e^{P(\alpha-1)\tau^*(t_2,y)}m(\tau^*(t_2,y))\Big]\bigg\} \le 0
\end{aligned}$$

since $\tau_1 > \tau^*(t_2,y)$, $E^Q\{(e^{-\bar{\mu}\tau_1} - e^{-\bar{\mu}\tau^*(t_2,y)})b_o\} \le 0$, and $e^{P(\alpha-1)\,r}\,m(r)$ is a Q-supermartingale. The proof still holds if $t_2 = T$ since then $\tau^*(t_2,y) = 0$ and $\tau_1 = T - t_1$, so that $\mathbf{1}_{\{t_1+\tau_1<T\}} = \mathbf{1}_{\{t_2<T\}} = 0$.

We now prove $[\mathrm{vii}]_{\hat{v}}$. For $t_1 < t_2$ in $[0,T]$ and $\tau_2 := \tau^*(t_1,y) \wedge (T-t_2)$ we have

$$\begin{aligned}
0 &\le \hat{v}(t_2,y) - \hat{v}(t_1,y)\\
&\le E^Q\Big\{e^{-\bar{\mu}\tau_2}H(t_2+\tau_2, Y^y(\tau_2)) - e^{-\bar{\mu}\tau^*(t_1,y)}H(t_1+\tau^*(t_1,y), Y^y(\tau^*(t_1,y)))\Big\}\\
&\le E^Q\Big\{e^{-\bar{\mu}\tau_2}\varphi(t_2+\tau_2, Y^y(\tau_2)) - e^{-\bar{\mu}\tau^*(t_1,y)}\varphi(t_1+\tau^*(t_1,y), Y^y(\tau^*(t_1,y)))\Big\}
\end{aligned}$$

since for ω in $\{\tau^*(t_1,y) = T - t_1\}$ we have $\mathbf{1}_{\{t_1+\tau^*(t_1,y)<T\}} = \mathbf{1}_{\{t_2+\tau_2<T\}} = 0$, for ω in $\{T - t_2 \le \tau^*(t_1,y) < T - t_1\}$ we have $\mathbf{1}_{\{t_2+\tau_2<T\}} = 0$ whereas $-\mathbf{1}_{\{t_1+\tau^*(t_1,y)<T\}} \le 0$, and, finally, for ω in $\{\tau^*(t_1,y) < T - t_2\}$ we have $\tau_2 = \tau^*(t_1,y)$ and the two terms involving $\frac{1}{f_C}$ cancel.

Therefore,

$$\begin{aligned}
0 &\le \hat{v}(t_2,y) - \hat{v}(t_1,y)\\
&\le y^{\alpha-1}E^Q\bigg\{\frac{e^{-P(\alpha-1)(t_2+\tau_2)}-1}{-P(\alpha-1)}\,e^{P(\alpha-1)\,\tau_2}\,m(\tau_2)\\
&\qquad\qquad - \frac{e^{-P(\alpha-1)(t_1+\tau^*(t_1,y))}-1}{-P(\alpha-1)}\,e^{P(\alpha-1)\,\tau^*(t_1,y)}\,m(\tau^*(t_1,y))\bigg\},
\end{aligned}$$

but $e^{P(\alpha-1)\,r}\,m(r)$ is a Q-supermartingale and $\tau_2 \le \tau^*(t_1,y)$, and hence

$$\begin{aligned}
0 &\le \hat{v}(t_2,y) - \hat{v}(t_1,y)\\
&\le \frac{y^{\alpha-1}}{-P(\alpha-1)}E^Q\Big\{\Big[e^{-P(\alpha-1)\,t_2}m(\tau_2) - e^{-P(\alpha-1)\,t_1}m(\tau^*(t_1,y))\Big]\Big\}\\
&= \frac{y^{\alpha-1}}{-P(\alpha-1)}\Big[e^{-P(\alpha-1)\,t_2} - e^{-P(\alpha-1)\,t_1}\Big].
\end{aligned}$$

It follows that $\hat{v}$ is continuous in t uniformly in $(t,y) \in [0,T] \times [\varepsilon,\infty)$ for any $\varepsilon > 0$, and hence $\hat{v}$ is continuous in $(t,y) \in [0,T] \times (0,\infty)$; that is, $[\mathrm{viii}]_{\hat{v}}$ holds.

Finally, for $t_1 < t_2$ in $[0,T)$, the above inequality implies that $\hat{v}(t_2,y) - \hat{v}(t_1,y) \leq H(t_2,y) - H(t_1,y)$, and so $[\text{ix}]_{\hat{v}}$ is proved. □

We are now able to say more about the continuation region (cf. (4.6)).

Proposition 4.2. *The following properties hold for the continuation region* Δ*:*

$[\text{i}]_\Delta$ *for each* $t \in [0,T)$*, the* t*-section* Δ_t *of the continuation region is open in* $(0,\infty)$*;*

$[\text{ii}]_\Delta$ *the* t*-section* Δ_t *of the continuation region is a semi-infinite interval;*

$[\text{iii}]_\Delta$ $\Delta \cap \{t > 0\}$ *is an open set.*

Proof. Part $[\text{i}]_\Delta$ follows from the continuity in y of $\hat{v}$ and H. For $[\text{ii}]_\Delta$ it suffices to point out that $[\text{iv}]_{\hat{v}}$ implies that if $x > y$ and $y \in \Delta_t$, then $x \in \Delta_t$. On the other hand, $[\text{iii}]_\Delta$ follows from the continuity of $\hat{v}$ in (t,y). □

Now $[\text{ii}]_\Delta$ implies that Δ_t must have the form $(\hat{y}(t),\infty)$ for some nonnegative number $\hat{y}(t)$.

Proposition 4.3. *Let* $\hat{y}(t)$ *be the function representing the boundary of the continuation region* Δ*. Then the following hold:*

$[\text{i}]_{bdy}$ $\hat{y}(t)$ *is nonincreasing and left-continuous for* $t < T$*;*

$[\text{ii}]_{bdy}$ $\hat{y}(t)$ *satisfies* $0 < \hat{y}(t)$ *for* $t < T$*;*

$[\text{iii}]_{bdy}$ $\hat{y}(T-) = 0$.

Proof. For $t \in [0,T)$, $\varepsilon > 0$ small enough and for any $\delta > 0$ we have $0 > \hat{v}(t,\hat{y}(t)+\delta) - H(t,\hat{y}(t)+\delta) \geq \hat{v}(t+\varepsilon,\hat{y}(t)+\delta) - H(t+\varepsilon,\hat{y}(t)+\delta)$. Therefore, $(\hat{y}(t)+\delta)$ is in $\Delta_{t+\varepsilon}$, hence $\hat{y}(t+\varepsilon) \leq \hat{y}(t)+\delta$, but $\delta > 0$ is arbitrary, and we obtain $\hat{y}(t+\varepsilon) \leq \hat{y}(t)$.

Now let $t > 0$ and $\varepsilon > 0$ small enough; then $\hat{y}(t) \leq \hat{y}(t-\varepsilon)$ and hence $\hat{y}(t) \leq \lim_{\varepsilon\downarrow 0} \hat{y}(t-\varepsilon) := \hat{y}(t-)$. On the other hand, $(t-\varepsilon, \hat{y}(t-\varepsilon)) \in ([0,T]\times[0,\infty)) - (\Delta \cap \{t > 0\})$, which is a closed set; thus also $(t,\hat{y}(t-)) \in ([0,T]\times[0,\infty)) - (\Delta\cap\{t>0\})$. It now follows from the definition of $\hat{y}(t)$ that $\hat{y}(t-) \leq \hat{y}(t)$, and we conclude that $\hat{y}(t-) = \hat{y}(t)$. Hence $[\text{i}]_{bdy}$ is proved.

To show that $\hat{y}(t) > 0$ assume the contrary, i.e., $\hat{y}(t) = 0$ for $t < T$. Then $[\text{i}]_{bdy}$ implies that $\hat{y}(t+s) = 0$ for all $s > 0$. For $y \in \Delta_t$ it follows that $Y^y(s) > 0$ for any $s \in [0,T-t]$. Hence $\tau^*(t,y) = T-t$ and

$$\begin{aligned}\hat{v}(t,y) &= E^Q\Big\{e^{-\bar{\mu}(T-t)}\varphi(T,Y^y(T-t))\Big\}\\ &= y^{\alpha-1}\,\frac{e^{-P(\alpha-1)T}-1}{-P(\alpha-1)}e^{P(\alpha-1)(T-t)}\,E^Q\{m(T-t)\}\\ &= y^{\alpha-1}\left[\frac{e^{-P(\alpha-1)t}-e^{P(\alpha-1)(T-t)}}{-P(\alpha-1)}\right] \qquad \forall y > 0.\end{aligned}$$

However, $H(t,y) = \frac{e^{-P(\alpha-1)T}-1}{-P(\alpha-1)}\,y^{\alpha-1} + \frac{1}{f_C}$, so $(\hat{v}-H)(t,y) = y^{\alpha-1}\,[\frac{-e^{P(\alpha-1)(T-t)}+1}{-P(\alpha-1)}] - \frac{1}{f_C} > 0$ if y is small. This contradiction proves $[\text{ii}]_{bdy}$.

Notice that $\hat{y}(T-)$ exists due to $[\text{i}]_{bdy}$. If $\hat{y}(T-) > 0$, then $\lim_{y\downarrow\hat{y}(T-)}\hat{v}(T,y) = H(T,\hat{y}(T-)) = \varphi(T,\hat{y}(T-))$ and $\lim_{t\uparrow T}\hat{v}(t,\hat{y}(t)) = \lim_{t\uparrow T} H(t,\hat{y}(t)) = \varphi(T,\hat{y}(T-)) + \frac{1}{f_C}$. This contradicts the continuity of $\hat{v}$ on $[0,T]\times(0,\infty)$. □

We then have (cf. (4.6))

$$\Delta = \{(t,y) \in [0,T)\times(0,\infty) : y > \hat{y}(t)\} \tag{4.10}$$

and we set

$$\Delta_{left} := \{(t,y) \in [0,T)\times(0,\infty) : y < \hat{y}(t)\}.$$

THEOREM 4.4. *The value function $\hat{v}(t,y)$ of the optimal stopping problem* (4.4) *is the unique nonnegative solution of the following $\hat{y}$-optimality conditions:*

(1) $-\hat{v}_t(t,y) - (L-\bar{\mu})\hat{v}(t,y) = 0$, $(t,y) \in \Delta \cap \{t > 0\}$;

(2) $-\hat{v}_t(t,y) - (L-\bar{\mu})\hat{v}(t,y) \leq 0$, $(t,y) \in \Delta_{left} \cap \{t > 0\}$;

(3) $\hat{v}(t,y) < H(t,y)$, $(t,y) \in \Delta$;

(4) $\hat{v}(t,y) = H(t,y)$, $(t,y) \in \Delta_{left}$;

(5) $\hat{v}(T,y) = \varphi(T,y) = H(T,y)$, $y \in (0,\infty)$;

(6) $\hat{v}(t,\hat{y}(t)) = H(t,\hat{y}(t))$, $t \in [0,T)$.

Proof. The proof of (1)–(6) is based on standard arguments. (3), (4), and (5) hold by construction; (6) follows by continuity in y.

To show (1) let $(s,y) \in \Delta \cap \{t > 0\}$. For $\varepsilon > 0$ and $y_1 < y < y_2$ define the rectangle $\mathcal{R} = (s-\varepsilon, s+\varepsilon) \times (y_1,y_2)$ such that its closure $\mathrm{cl}(\mathcal{R}) \subset \Delta$. Define $\partial_\circ \mathcal{R} := \partial\mathcal{R} \setminus [\{s-\varepsilon\} \times (y_1,y_2)]$ of $\mathcal{R}$ and consider the initial-boundary value problem

$$\begin{cases} -f_t(t,y) - (L-\bar{\mu})f(t,y) = 0 & \text{in } \mathcal{R}, \\ f = \hat{v} \quad \text{on } \partial_\circ\mathcal{R}. \end{cases}$$

Then by reversing time, $t \to T-t$, this problem corresponds to a classical initial value problem with uniformly parabolic operator in $\mathcal{R}$ (due to $s+\varepsilon < T$ and Proposition 4.3) and parabolic boundary $\partial_\circ\mathcal{R}$. Hence the classical theory may be applied to obtain the unique solution f with f_t, f_y, f_{yy} continuous. It remains to show that f coincides with $\hat{v}$ in $\mathcal{R}$. Define the stopping time $\theta := \inf\{t \in [0,\varepsilon] : (s+t, Y^y(t)) \in \partial_\circ\mathcal{R}\}$ and the process $F(t) := e^{-\bar{\mu}t} f(s+t, Y^y(t))$; then

$$\begin{aligned} F(0) &= f(s,y) \\ &= E^Q\left\{F(\theta) + \int_0^\theta e^{-\bar{\mu}t}\Big[-f_t(s+t,Y^y(t)) - (L-\bar{\mu})f(s+t,Y^y(t))\Big]dt + \text{martingale}\right\}; \end{aligned}$$

that is, $f(s,y) = E^Q\{F(\theta)\} = E^Q\{e^{-\bar{\mu}\theta}\hat{v}(s+\theta, Y^y(\theta))\}$. Recall that $\tau^*(s,y) = \inf\{t \in [0,T-s] : \hat{v}(s+t,Y^y(t)) = H(s+t,Y^y(t))\}$; then the Markov property and the fact that Q-a.s. $\theta \leq \tau^*(s,y)$, the optimal stopping time for (s,y), imply that $\tau^*(s,y) = \theta + \tau(\theta)$ a.s., where we have set $\tau(\theta) := \tau^*(s+\theta, Y^y(\theta))$. Therefore, $E^Q\{e^{-\bar{\mu}\theta}\hat{v}(s+\theta,Y^y(\theta))\} = E^Q\{e^{-\bar{\mu}\theta}e^{-\bar{\mu}\tau(\theta)}H(s+\theta+\tau(\theta), Y^{Y^y(\theta)}(\tau(\theta)))\}$, hence $f(s,y) = E^Q\{e^{-\bar{\mu}\theta}\hat{v}(s+\theta,Y^y(\theta))\} = \hat{v}(s,y)$, and (1) follows.

For (2) we proceed by contradiction. If at some $(s,x) \in \Delta_{left} \cap \{t > 0\}$ we had $-\hat{v}_t(s,y) - (L-\bar{\mu})\hat{v}(s,y) > 0$, then the same inequality would hold in a ball $B \subset \Delta_{left} \cap \{t > 0\}$ centered in (s,y), thanks to (4) and the smoothness of H. Note that $\hat{y}(s) > 0$ since $s < T$. Hence the Itô formula applied to $e^{-\bar{\mu}t}H(s+t,Y^y(t))$ up to the first exit time τ_B of $(s+t,Y^y(t))$ from B would give $H(s,y) > E^Q\{e^{-\bar{\mu}\tau_B}H(s+\tau_B, Y^y(\tau_B))\}$, but $\hat{v}(s,y) = H(s,y)$ (again by point (4)), and we would contradict the definition of $\hat{v}$.

As for the uniqueness, let $\psi(t,y)$ be another nonnegative solution of (1)–(6) (with Δ given by $\hat{y}$). For $t \in [0,T)$ define $\Delta^\varepsilon := \{(t,y) \in \Delta : \|(T-t,y)\| > \varepsilon\}$. For $(t,y) \in \Delta^\varepsilon$ set $\tau^\varepsilon(t,y) := \inf\{s > 0 : (t+s, Y^y(s)) \notin \Delta^\varepsilon\}$. Note that $\tau^\varepsilon(t,y) \uparrow \tau^*(t,y) = \inf\{s \in [0,T-t] : Y^y(s) = \hat{y}(t+s)\} \wedge (T-t)$ a.s. as $\varepsilon \downarrow 0$ since $Y^y(\tau^*(t,y)) \neq 0$ a.s. It follows from (1)–(6) that the process $G^\varepsilon(s) := e^{-\bar{\mu}(s\wedge\tau^\varepsilon(t,y))}\psi(t+(s\wedge\tau^\varepsilon(t,y)), Y^y(s\wedge\tau^\varepsilon(t,y)))$ is a bounded martingale; hence by using (5) and (6) we obtain

$$\begin{aligned} \psi(t,y) = G^\varepsilon(0) = E^Q\{G^\varepsilon(T-t)\} &= E^Q\{e^{-\bar{\mu}\tau^\varepsilon(t,y)}\psi(t+\tau^\varepsilon(t,y), Y^y(\tau^\varepsilon(t,y)))\} \\ &= E^Q\{e^{-\bar{\mu}\tau^\varepsilon(t,y)}H(t+\tau^\varepsilon(t,y), Y^y(\tau^\varepsilon(t,y)))\,\mathbf{1}_{\{\tau^\varepsilon(t,y)=\tau^*(t,y)\}}\} \end{aligned}$$

$$
\begin{aligned}
&+ E^Q\{e^{-\bar{\mu}\tau^\varepsilon(t,y)}\psi(t+\tau^\varepsilon(t,y), Y^y(\tau^\varepsilon(t,y)))\,\mathbf{1}_{\{\tau^\varepsilon(t,y)<\tau^*(t,y)\}}\} \\
&= \hat{v}(t,y) + E^Q\Big\{\Big[e^{-\bar{\mu}\tau^\varepsilon(t,y)}\psi(t+\tau^\varepsilon(t,y), Y^y(\tau^\varepsilon(t,y))) \\
&\qquad - e^{-\bar{\mu}\tau^*(t,y)}H(t+\tau^*(t,y), Y^y(\tau^*(t,y)))\Big]\,\mathbf{1}_{\{\tau^\varepsilon(t,y)<\tau^*(t,y)\}}\Big\},
\end{aligned}
$$

since $\tau^*(t,y)$ is optimal for $\hat{v}(t,y)$. Observe that $\mathbf{1}_{\{\tau^\varepsilon(t,y)<\tau^*(t,y)\}} \downarrow 0$ a.s. as $\varepsilon \downarrow 0$. Moreover, by (4.1) for $p > 1$

$$
\begin{aligned}
&E^Q\Big\{\Big[e^{-\bar{\mu}\tau^\varepsilon(t,y)}\psi(t+\tau^\varepsilon(t,y), Y^y(\tau^\varepsilon(t,y)))\Big]^p\Big\} \\
&\le E^Q\Big\{\Big[e^{-\bar{\mu}\tau^\varepsilon(t,y)}H(s+\tau^\varepsilon(t,y), Y^y(\tau^\varepsilon(t,y)))\Big]^p\Big\} \\
&\le k_p E^Q\Big\{\Big[\frac{e^{-P(\alpha-1)t} - e^{P(\alpha-1)\,\tau^\varepsilon(t,y)}}{-P(\alpha-1)}\, m(\tau^\varepsilon(t,y))\, y^{\alpha-1}\Big]^p + \frac{1}{f_C^p}\Big\} \\
&\le K_p\Big[\Big(\frac{e^{-P(\alpha-1)t}}{-P(\alpha-1)}\Big)^p y^{(\alpha-1)p} + 1\Big] < \infty
\end{aligned}
$$

since $E^Q\{[m(\tau^\varepsilon(t,y))]^p\} = e^{\frac{1}{2}p(p-1)(\alpha-1)^2\,\|\sigma_C\|^2\,\tau^\varepsilon(t,y)} \le e^{\frac{1}{2}p(p-1)(\alpha-1)^2\,\|\sigma_C\|^2\,(T-t)} < \infty$. Similarly $E^Q\{[e^{-\bar{\mu}\tau^*(t,y)}H(t+\tau^*(t,y), Y^y(\tau^*(t,y)))]^p\} < \infty$. Hence by uniform integrability $\psi = \hat{v}$ on Δ^ε for any $\varepsilon > 0$, that is, on Δ. □

COROLLARY 4.5. *The function $\hat{y}(t)$ representing the boundary of the continuation region Δ is bounded above; in fact,*

[iv]$_{bdy}$ $\hat{y}(t) \le y_\circ := (\frac{\bar{\mu}}{f_C})^{\frac{1}{\alpha-1}}$ *for $t < T$.*

Proof. Set $U := \{(t,y) \in [0,T) \times (0,\infty) : -H_t(t,y) - (L-\bar{\mu})H(t,y) > 0\}$. Since $H = \varphi + \frac{1}{f_C}$ on Δ, Itô's formula implies that $H(t,y) > E^Q\{e^{-\bar{\mu}\,\tau^*(t,y)}[\varphi(t+\tau^*(t,y), Y^y(\tau^*(t,y))) + \frac{1}{f_C}]\} \ge E^Q\{e^{-\bar{\mu}\,\tau^*(t,y)}[H(t+\tau^*(t,y), Y^y(\tau^*(t,y)))]\} \ge \hat{v}(t,y)$ for $(t,y) \in \Delta$. Hence $U \subset \Delta$; that is, it is never optimal to stop before the process exits from U, i.e., when $-y^{\alpha-1} + \frac{\bar{\mu}}{f_C}\mathbf{1}_{\{t<T\}} > 0$. □

We shall now show that the free boundary may be characterized as the unique solution of an integral equation in the spirit of [13]. Most of the arguments below are similar to those used by Jacka [13] for the free boundary of the American put or, more generally, by Pedersen and Peskir [17]. However, these arguments require the smooth fit property, i.e.,

$$
\hat{v}_y(t,\hat{y}(t)) = H_y(t,\hat{y}(t)), \qquad t \in [0,T). \tag{4.11}
$$

The concavity property [ii]$_{\hat{v}}$ (cf. Proposition 4.1) combined with the differentiability of $y^{\frac{1}{\alpha-1}}$ guarantees that the one-sided y-derivatives of $\hat{v}$ are defined and satisfy $\hat{v}_y(t,\hat{y}(t)-) = H_y(t,\hat{y}(t)) \ge \hat{v}_y(t,\hat{y}(t)+)$ since $\hat{v}(t,\cdot) = H(t,\cdot)$ on the complement of Δ_t and $(\hat{v} - H)(t,y)$ is nonincreasing in y for $t < T$. For smooth fit it remains to show that $\hat{v}_y(t,\hat{y}(t)+) \ge H_y(t,\hat{y}(t))$. Since our $\hat{v}$ is an inf rather than a sup as in the option problem we are unable to show this without further assumptions.

Assumption-[Cfb].

$$
\text{The free boundary } \hat{y}(t) \text{ is continuous on } [0,T). \tag{4.12}
$$

LEMMA 4.6. *Under Assumption-*[Cfb] *the smooth fit property* (4.11) *holds.*

Proof. We show that $\hat{v}_y(t,\hat{y}(t)+) \geq H_y(t,\hat{y}(t))$. Fix $t<T$ and set $\hat{y} := \hat{y}(t)$; then for $\varepsilon > 0$ the inf in $\hat{v}(t,\hat{y}+\varepsilon)$ is attained at $\tau^*_\varepsilon := \tau^*(t,\hat{y}+\varepsilon)$ (cf. (4.7)), and this is nondecreasing in ε. So $\hat{v}(t,\hat{y}+\varepsilon)$ is equal to $E^Q\{e^{-\bar{\mu}\tau^*_\varepsilon}H(t+\tau^*_\varepsilon, Y^{\hat{y}+\varepsilon}(\tau^*_\varepsilon))\}$; now by subtracting $\hat{v}(t,\hat{y})$ (a quantity less than or equal to $E^Q\{e^{-\bar{\mu}\tau^*_\varepsilon}H(t+\tau^*_\varepsilon, Y^{\hat{y}}(\tau^*_\varepsilon))\}$) we obtain $\hat{v}(t,\hat{y}+\varepsilon)-\hat{v}(t,\hat{y}) \geq E^Q\{e^{-\bar{\mu}\tau^*_\varepsilon}[H(t+\tau^*_\varepsilon, Y^{\hat{y}+\varepsilon}(\tau^*_\varepsilon)) - H(t+\tau^*_\varepsilon, Y^{\hat{y}}(\tau^*_\varepsilon))]\}$. Hence if we divide by ε and take limits as $\varepsilon \downarrow 0$, then we get (cf. (4.1) and (4.2))

$$\hat{v}_y(t,\hat{y}+) \geq \lim_{\varepsilon\downarrow 0} \frac{[(\hat{y}+\varepsilon)^{\alpha-1}-(\hat{y})^{\alpha-1}]}{\varepsilon} E^Q\left\{\frac{e^{-P(\alpha-1)(t+\tau^*_\varepsilon)}-1}{-P(\alpha-1)} e^{P(\alpha-1)\,\tau^*_\varepsilon} m(\tau^*_\varepsilon)\right\}.$$

Notice that $\tau^*_\varepsilon \downarrow \tau_o$ as $\varepsilon \downarrow 0$, where τ_o is the first time $(t+\cdot, Y^{\hat{y}}(\cdot))$ hits the boundary of $\{(s,y)\in\Delta : s>t\}$. Now $\hat{y}>0$ implies that the diffusion $Y^{\hat{y}}$ is nondegenerate and $\hat{y}(t)=\hat{y}(t+)$ by assumption; therefore $\tau_o = 0$. So $\frac{e^{-P(\alpha-1)(t+\tau^*_\varepsilon)}-1}{-P(\alpha-1)} e^{P(\alpha-1)\,\tau^*_\varepsilon} m(\tau^*_\varepsilon)$ converges to $\frac{e^{-P(\alpha-1)t}-1}{-P(\alpha-1)}$ a.s. Also,

$$\left|\frac{e^{-P(\alpha-1)(t+\tau^*_\varepsilon)}-1}{-P(\alpha-1)} e^{P(\alpha-1)\,\tau^*_\varepsilon}\right| \leq K_1,$$

where K_1 is a constant depending on $T-t$ since $\tau^*_\varepsilon \in [0,T-t]$, and standard inequalities imply that $E^Q\{(m(\tau^*_\varepsilon)^p\} \leq e^{\frac{p(p-1)}{2}(\alpha-1)^2\|\sigma_C\|^2(T-t)}$. Hence

$$\frac{e^{-P(\alpha-1)(t+\tau^*_\varepsilon)}-1}{-P(\alpha-1)} e^{P(\alpha-1)\,\tau^*_\varepsilon} m(\tau^*_\varepsilon)$$

is uniformly integrable. It follows that

$$\hat{v}_y(t,\hat{y}+) \geq (\alpha-1)\hat{y}^{\alpha-2}\frac{e^{-P(\alpha-1)t}-1}{-P(\alpha-1)} = H_y(t,\hat{y}). \qquad \square$$

We point out that if the free boundary is continuous, then, contrary to what we claimed in [6], the smooth fit condition (4.11) holds also for the original v since $v = \hat{v} - \varphi$. There are examples of solutions to optimal stopping problems where smooth fit fails, e.g., [16]. In our case the regularity of the data gives no indication that Assumption-[Cfb] should fail; hence we assume [Cfb], although we are not able to prove it.

LEMMA 4.7. *Under Assumption-*[Cfb] *the process*

(4.13) $M_s(t)$

$$:= e^{-\bar{\mu}t}\hat{v}(s+t, Y^y(t)) + \int_0^t e^{-\bar{\mu}r}\left[-(Y^y(r))^{\alpha-1} + \frac{\bar{\mu}}{f_C}\mathbf{1}_{\{s+r<T\}}\right]\mathbf{1}_{\{Y^y(r)\leq\hat{y}(s+r)\}}dr,$$

$t \in [0,T-s]$, *is a martingale for* $s\in(0,T)$ *and* $y>0$.

Proof. It follows from (1) and (4) of Theorem 4.4 that $-\hat{v}_t(t,y) - (L-\bar{\mu})\hat{v}(t,y)$ equals $-y^{\alpha-1} + \frac{\bar{\mu}}{f_C}\mathbf{1}_{\{t<T\}}$ if $y<\hat{y}(t)$ and $t\in(0,T)$, whereas it equals zero for $y>\hat{y}(t)$ and $t\in(0,T)$. By Lemma 4.6, $\hat{v}$ is C^1 in y due to the smooth fit property, so $dM_s(t) = e^{-\bar{\mu}t}\hat{v}_y(s+t,Y^y(t))\,Y^y(t)\,\|\sigma_C\|\,dB^Q(t)$ and

$$E^Q\left\{\int_0^{T-s}\left|e^{-\bar{\mu}t}\hat{v}_y(s+t,Y^y(t))\,Y^y(t)\,\|\sigma_C\|\right|^2 dt\right\}$$

$$\leq \int_0^{T-s} E^Q\Bigg\{e^{-2\bar{\mu}t}\|\sigma_C\|^2\left[\frac{\alpha-1}{-P(\alpha-1)}\right]^2 e^{-2P(\alpha-1)t}\, y^{2\alpha-1}$$

$$\times\, e^{2(\alpha-1)\|\sigma_C\|B^Q(t)+2(\alpha-1)(\frac{1}{2}\|\sigma_C\|^2-\mu_C)t}\Bigg\}dt$$

$$= y^{2\alpha-1}\|\sigma_C\|^2\left[\frac{\alpha-1}{-P(\alpha-1)}\right]^2$$

$$\times \int_0^{T-s} e^{(\alpha-1)^2\|\sigma_C\|^2 t} E^Q\Big\{e^{2(\alpha-1)\|\sigma_C\|B^Q(t)-\frac{1}{2}(4(\alpha-1)^2\|\sigma_C\|^2 t}\Big\}dt$$

$$= y^{2\alpha-1}\|\sigma_C\|^2\left[\frac{\alpha-1}{-P(\alpha-1)}\right]^2 \int_0^{T-s} e^{(\alpha-1)^2\|\sigma_C\|^2 t}dt$$

$$= y^{2\alpha-1}\,\frac{e^{(\alpha-1)^2\|\sigma_C\|^2(T-s)}-1}{[-P(\alpha-1)]^2} < \infty$$

due to the estimate of $\hat{v}_y$ in Proposition 4.1, $[\mathrm{v}]_{\hat{v}}$. Hence M_s is indeed a martingale. □

We now have the following theorem.

THEOREM 4.8. *Under Assumption-*[Cfb] *the free boundary* $\hat{y}(\cdot)$ *is the unique left-continuous solution* $h(\cdot)$ *of the integral equation*

$$(4.14)\quad H(s,y) = E^Q\Bigg\{e^{-\bar{\mu}(T-s)}H(T,Y^y(T-s))$$

$$+\int_0^{T-s} e^{-\bar{\mu}t}\left[-(Y^y(t))^{\alpha-1}+\frac{\bar{\mu}}{f_C}\,\mathbf{1}_{\{s+t<T\}}\right]\mathbf{1}_{\{Y^y(t)\leq h(s+t)\}}\,dt\Bigg\}\qquad \forall y\leq h(s),$$

satisfying $0<h(s)\leq y_\circ$ *for all* $s<T$.

Proof. As $\hat{v}(T,y)=H(T,y)$ for any y, then for $h=\hat{y}$ Lemma 4.7 gives $H(s,y)=E^Q M_s(T-s)$. Since $M_s(0)=\hat{v}(s,y)=H(s,y)$ for $y\leq\hat{y}(s)$, the martingale property of $M_s(t)$ implies that $\hat{y}$ is a solution of the integral equation above.

Now let $h(t)\leq y_o$ be another left-continuous solution and set

$$\phi(s,y) := E^Q\Bigg\{e^{-\bar{\mu}(T-s)}H(T,Y^y(T-s))$$

$$+\int_0^{T-s} e^{-\bar{\mu}r}\left[-(Y^y(r))^{\alpha-1}+\frac{\bar{\mu}}{f_C}\,\mathbf{1}_{\{s+r<T\}}\right]\mathbf{1}_{\{Y^y(r)\leq h(s+r)\}}\,dr\Bigg\}.$$

Then $\phi(T,y)=H(T,y)$. For fixed $s\in(0,T)$, $t\in[0,T-s]$, and $y>0$,

$$e^{-\bar{\mu}t}\phi(s+t,Y^y(t)) = E^Q\Bigg\{e^{-\bar{\mu}(T-s)}H(T,Y^z(T-s-t))$$

$$+\int_t^{T-s} e^{-\bar{\mu}r}\left[-(Y^z(r-t))^{\alpha-1}+\frac{\bar{\mu}}{f_C}\,\mathbf{1}_{\{s+r<T\}}\right]\mathbf{1}_{\{Y^z(r-t)\leq h(s+r)\}}\,dr\Bigg\}\Bigg|_{z=Y^y(t)}$$

$$= E^Q\Bigg\{e^{-\bar{\mu}(T-s)}H(T,Y^y(T-s))$$

$$+\int_t^{T-s} e^{-\bar{\mu}r}\left[-(Y^y(r))^{\alpha-1}+\frac{\bar{\mu}}{f_C}\mathbf{1}_{\{s+r<T\}}\right]\mathbf{1}_{\{Y^y(r)\le h(s+r)\}}\,dr\bigg|\mathcal{F}_t\bigg\},$$

and hence the process

$$\begin{aligned}N_s(t) &:= e^{-\bar{\mu}t}\phi(s+t,Y^y(t))\\ &\qquad+\int_0^t e^{-\bar{\mu}r}\left[-(Y^y(r))^{\alpha-1}+\frac{\bar{\mu}}{f_C}\mathbf{1}_{\{s+r<T\}}\right]\mathbf{1}_{\{Y^y(r)\le h(s+r)\}}dr\\ &= E^Q\bigg\{e^{-\bar{\mu}(T-s)}H(T,Y^y(T-s))\\ &\qquad+\int_0^{T-s} e^{-\bar{\mu}r}\left[-(Y^y(r))^{\alpha-1}+\frac{\bar{\mu}}{f_C}\mathbf{1}_{\{s+r<T\}}\right]\mathbf{1}_{\{Y^y(r)\le h(s+r)\}}\,dr\bigg|\mathcal{F}_t\bigg\}\end{aligned}$$

is a martingale on $[0,T-s]$. As h is a solution of (4.14), $\phi(s,y)=H(s,y)$ for $y\le h(s)$. The proof now follows along the same lines as that of [13, Theorem 4.2.2].

Pick $y>h(s)$ and define the stopping time $\tau:=\inf\{t\in[0,T-s):Y^y(t)\le h(s+t)\}\wedge(T-s)$; then $\phi(s,y)=N_s(0)=E^Q\{e^{-\bar{\mu}\tau}\phi(s+\tau,Y^y(\tau))\}=E^Q\{e^{-\bar{\mu}\tau}H(s+\tau,Y^y(\tau))\}$, and hence $\phi(s,y)\ge\hat{v}(s,y)$. On the other hand, for $y\le h(s)$, we have $\phi(s,y)=H(s,y)\ge\hat{v}(s,y)$. Therefore, $\phi(s,y)\ge\hat{v}(s,y)$ for all $s\in(0,T)$ and $y>0$.

Now pick $0<y\le h(s)\wedge\hat{y}(s)$ (so $\hat{v}(s,y)=H(s,y)=\phi(s,y)$) and define the stopping time $\tau':=\inf\{t\in[0,T-s):Y^y(t)\ge\hat{y}(s+t)\}\wedge(T-s)$; then

$$\begin{aligned}0&=\hat{v}(s,y)-\phi(s,y)=M_s(0)-N_s(0)\\ &=E^Q\Big\{e^{-\bar{\mu}\tau'}\Big[\hat{v}(s+\tau',Y^y(\tau'))-\phi(s+\tau',Y^y(\tau'))\Big]\Big\}\\ &\quad+E^Q\bigg\{\int_0^{\tau'} e^{-\bar{\mu}r}\left[-(Y^y(r))^{\alpha-1}+\frac{\bar{\mu}}{f_C}\mathbf{1}_{\{s+r<T\}}\right]\\ &\qquad\times\Big[\mathbf{1}_{\{Y^y(r)\le\hat{y}(s+r)\}}-\mathbf{1}_{\{Y^y(r)\le h(s+r)\}}\Big]dr\bigg\},\end{aligned}$$

and $Y^y(\cdot)\le\hat{y}(s+\cdot)$ on $[0,\tau')$ implies that

$$\begin{aligned}0&=E^Q\Big\{e^{-\bar{\mu}\tau'}\Big[\hat{v}(s+\tau',Y^y(\tau'))-\phi(s+\tau',Y^y(\tau'))\Big]\Big\}\\ &\quad+E^Q\bigg\{\int_0^{\tau'} e^{-\bar{\mu}r}\left[-(Y^y(r))^{\alpha-1}+\frac{\bar{\mu}}{f_C}\mathbf{1}_{\{s+r<T\}}\right]\mathbf{1}_{\{Y^y(r)>h(s+r)\}}dr\bigg\}.\end{aligned}\tag{4.15}$$

Since both terms on the right-hand side of (4.15) are nonpositive (recall that $\hat{y}$ is below $y_\circ:=(\frac{\bar{\mu}}{f_C})^{\frac{1}{\alpha-1}}$ and hence $Y^y(\cdot)\le\hat{y}(s+\cdot)\le(\frac{\bar{\mu}}{f_C})^{\frac{1}{\alpha-1}}$ on $[0,\tau')$), it follows that both terms are zero. In particular, from the integral term we deduce that the Lebesgue $\times\, Q$ measure of $\{h(s+r)<Y^y(r)\le\hat{y}(s+r)\}\cap\{r<\tau'\}$ is zero for all s,y. Since the distribution of $Y^y(r)$ has support $(0,\infty)$, we conclude that $h\ge\hat{y}$ a.e. This fact, the assumption $h(\cdot)\le y_o$, and $H(T,Y^y(T-s))=\hat{v}(T,Y^y(T-s))$ in the definition of ϕ imply that $\phi(s,y)\le E^QM_s(T-s)=M_s(0)=\hat{v}(s,y)$, and hence $\phi=\hat{v}$

on $y > 0$. Define $\tau'' := \inf\{t \in [0, T-s) : Y^y(t) \geq h(s+t)\} \wedge (T-s)$, and proceed as for τ' to obtain

$$\begin{aligned} 0 = {} & E^Q\Big\{e^{-\bar{\mu}\tau''}\Big[\hat{v}(s+\tau'', Y^y(\tau'')) - \phi(s+\tau'', Y^y(\tau''))\Big]\Big\} \\ & - E^Q\bigg\{\int_0^{\tau''} e^{-\bar{\mu}r}\bigg[-(Y^y(r))^{\alpha-1} + \frac{\bar{\mu}}{f_C}\mathbb{1}_{\{s+r<T\}}\bigg]\mathbb{1}_{\{Y^y(r)>\hat{y}(s+r)\}}dr\bigg\}. \end{aligned}$$

As above, it follows that $\hat{y} \geq h$, hence $h = \hat{y}$ almost everywhere, and so $h \equiv \hat{y}$ by left continuity. □

LEMMA 4.9. *Under Assumption-*[Cfb] *for the free boundary the following representation holds:*

$$\begin{aligned} &\int_0^{T-s} e^{-\bar{\mu}t}E^Q\bigg\{\bigg[(Y^{\hat{y}(s)}(t))^{\alpha-1} - \frac{\bar{\mu}}{f_C}\mathbb{1}_{\{s+t<T\}}\bigg]\mathbb{1}_{\{Y^{\hat{y}(s)}(t)<\hat{y}(s+t)\}}\bigg\}\,dt \\ &\qquad = \bigg[\frac{e^{P(\alpha-1)(T-s)}-1}{P(\alpha-1)}\bigg](\hat{y}(s))^{\alpha-1} - \frac{1}{f_C}\mathbb{1}_{\{s<T\}}. \end{aligned} \tag{4.16}$$

Proof. For $y = \hat{y}(s)$ we have

$$e^{-\bar{\mu}(T-s)}E^Q\{H(T, Y^{\hat{y}(s)}(T-s))\} - H(s, \hat{y}(s)) = \bigg[\frac{e^{P(\alpha-1)(T-s)}-1}{P(\alpha-1)}\bigg]y^{\alpha-1} - \frac{1}{f_C}\mathbb{1}_{\{s<T\}}.$$

Then (4.14) implies (4.16). □

THEOREM 4.10. *Under Assumption-*[Cfb] *the optimal free boundary* $\hat{y}(\cdot)$ *is identified as the unique solution of the instantaneous-stopping equation*

$$\begin{aligned} &\Big(\hat{y}(s)\Big)^{\alpha-1}\int_0^{T-s} e^{P(\alpha-1)t}\Big[1 - \Phi\Big(\Psi(t; s, \hat{y}(s)) - (\alpha-1)\|\sigma_C\|\sqrt{t}\Big)\Big]dt \\ &\qquad + \frac{\bar{\mu}}{f_C}\int_0^{T-s} e^{-\bar{\mu}t}\,\Phi(\Psi(t; s, \hat{y}(s)))dt = \frac{1}{f_C}\mathbb{1}_{\{s<T\}}, \end{aligned} \tag{4.17}$$

where $\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}}e^{-r^2/2}dr$ *is the cumulative of the* $N(0,1)$*-distribution and*

$$\Psi(t; s, y) = \frac{1}{\|\sigma_C\|\sqrt{t}}\bigg[\ln\bigg(\frac{\hat{y}(s+t)}{y}\bigg) - \bigg(\frac{1}{2}\|\sigma_C\|^2 - \mu_C\bigg)\,t\bigg].$$

Proof. We calculate the expected value in (4.16), $E^Q\{\mathbb{1}_{\{Y^y(t)<\hat{y}(s+t)\}}\} = Q\{\omega \in \Omega : \|\sigma_C\|B^Q(\omega, t) + (\frac{1}{2}\|\sigma_C\|^2 - \mu_C)\,t < \ln(\frac{\hat{y}(s+t)}{y})\} = \Phi(\Psi(t; s, y))$. Similarly,

$$\begin{aligned} &E^Q\Big\{e^{-\bar{\mu}t}(Y^y(t))^{\alpha-1}\,\mathbb{1}_{\{Y^y(t)<\hat{y}(s+t)\}}\Big\} \\ &\quad = E^Q\Big\{y^{\alpha-1}\,e^{P(\alpha-1)t}\,m(t)\,\mathbb{1}_{\{\|\sigma_C\|\,B^Q(t)+(\frac{1}{2}\|\sigma_C\|^2-\mu_C)\,t<\ln(\frac{\hat{y}(s+t)}{y})\}}\Big\} \\ &\quad = y^{\alpha-1}\,e^{P(\alpha-1)t} \\ &\qquad \times E^Q\Big\{m(t)\mathbb{1}_{\{\|\sigma_C\|\,[B^Q(t)-(\alpha-1)\|\sigma_C\|\,t]<\ln(\frac{\hat{y}(s+t)}{y})-(\frac{1}{2}\|\sigma_C\|^2-\mu_C)\,t-(\alpha-1)\|\sigma_C\|^2\,t\}}\Big\} \\ &\quad = y^{\alpha-1}\,e^{P(\alpha-1)t}\,\Phi\Big(\Psi(t; s, y) - (\alpha-1)\|\sigma_C\|\sqrt{t}\Big). \qquad □ \end{aligned}$$

Having found $\hat{y}(t)$ from Theorem 4.10 we can find $\hat{v}(t,y)$ as the solution of the corresponding Cauchy problem in the continuation region Δ and as $H(t,y)$ elsewhere.

Now $V(t,y)$ (cf. (2.7)) is obtained as follows. Recall that the optimal investment is zero in the continuation region, i.e., until the first time the diffusion reaches the free boundary; then the dynamic programming equation and Dynkin's formula imply that $-V_t - \frac{1}{2}\|\sigma_C\|^2 y^2 V_{yy} + \mu_C\, y\, V_y + \mu_F V = y^\alpha/\alpha$ in Δ. Since $v(t,y) = \hat{v}(t,y) - \varphi(t,y) = (\hat{v} - H)(t,y) + \frac{1}{f_C}\mathbf{1}_{\{t<T\}}$ by (4.5) and the definition of H, and $V(t,y) = \int v(t,y)\, dy + h(t)$ for some function h with $h(T) = a_o$ by (3.11), we can substitute this representation into the above PDE. However, $\varphi_t(t,y) + (L - \bar{\mu})\varphi(t,y) = y^{\alpha-1}$ by construction and $-\hat{v}_t(t,y) - (L - \bar{\mu})\hat{v}(t,y) = 0$ in Δ by Theorem 4.4; hence we obtain $h_t(t) - \mu_F\, h(t) = 0$, which together with $h(T) = a_o$ gives $h(t) = a_o\, e^{-\mu_F(T-t)}$. Therefore, $V(t,y) = \int(\hat{v}(t,y) - \varphi(t,y))\, dy + a_o\, e^{-\mu_F(T-t)}$ for $y > \hat{y}(t)$.

On the other hand, for $y \leq \hat{y}(t)$ we have $V(t,y) = V(t,\hat{y}(t)) - \frac{\hat{y}(t)-y}{f_C}$ since $V_y = v = \frac{1}{f_C}$ here. In the above procedure $\hat{y}$ and $\hat{v}$ will usually have to be found numerically.

In [6, section 5] (cf. also [7]) we calculated a curve $y^*(t)$ which was incorrectly identified as the free boundary. A discrete approximation of the integral equation (4.17) allows us to compare the free boundary $\hat{y}(t)$ to $y^*(t)$. See Figure 1, where $\hat{y}$ is denoted by y.

5. Appendix. Recall from section 2 that the optimal L is $I^{u_A(\cdot,C)}(w)$ with $u_A(\cdot,C) := R^A(C,\cdot)$. We are now writing the C-dependence explicitly. We wish to present the next result in some generality as it is required in sections 2 and 3 as well as in [8]; the notation also corresponds to that used in [9] to which we will refer. Here x replaces L, y replaces w, $u(x,C)$ replaces $R(C,x)$, and $\hat{x}(y,C)$ replaces $I^{u_A(\cdot,C)}(w)$. The parameter μ is 0 here; cf. [9]. A is a general convex polyhedral set satisfying some assumptions (cf. the proposition below); these are satisfied by our current $A = [0,\kappa_L]$. We write $\mathcal{T} = [0,\infty)$ for the domain of the parameter C and $\mathcal{N}_A(x)$ for the cone of outward normals to A at x. Recall that the concave conjugate function of $u_A(\cdot,C)$ is

$$u_A^*(y,C) := \inf_{x\in A}\{x^\top y - u(x,C)\} = -\sup_{x\in A}\{u(x,C) - y^\top x\}. \tag{5.1}$$

As a conjugate function, $u_A^*(y,C)$ is concave in y for fixed C.

PROPOSITION 5.1. *Assume that A is polyhedral and that Assumptions* 1, 2, 3 *of* [9] *hold with the same μ for all C. If $u(\cdot\,,\cdot\,)$ is strictly concave and twice continuously differentiable, then*

$$\frac{\partial}{\partial C}u_A^*(y,C) = -\frac{\partial}{\partial C}u(x,C)\Big|_{x=\hat{x}(y,C)} \quad a.e.\ in\ C, \tag{5.2}$$

and $C \mapsto u_A^*(y,C)$ is strictly convex on $\mathcal{T}$ for each y where it is finite.

Proof. According to [19, Theorem 23.5(a*)] and [9, Proposition 3.2], the inf in (5.1) is attained at $\hat{x}(y,C) = I^{u_A(\cdot,C)}(y)$. Observe that $\hat{x}(y,C)$ is continuous and its C-derivative is bounded (cf. (3.9) of [9]); hence it is Lipschitz in C and $\frac{\partial}{\partial C}\hat{x}(y,C)$ exists a.e. For fixed y

$$\begin{aligned} -\frac{\partial}{\partial C}u_A^*(y,C) &= \frac{d}{dC}[u(\hat{x}(y,C),C) - y^\top \hat{x}(y,C)] \\ &= \frac{\partial}{\partial C}u(x,C)\Big|_{x=\hat{x}(y,C)} + [\nabla u(\hat{x}(y,C),C) - y]^\top \frac{\partial}{\partial C}\hat{x}(y,C) \end{aligned}$$

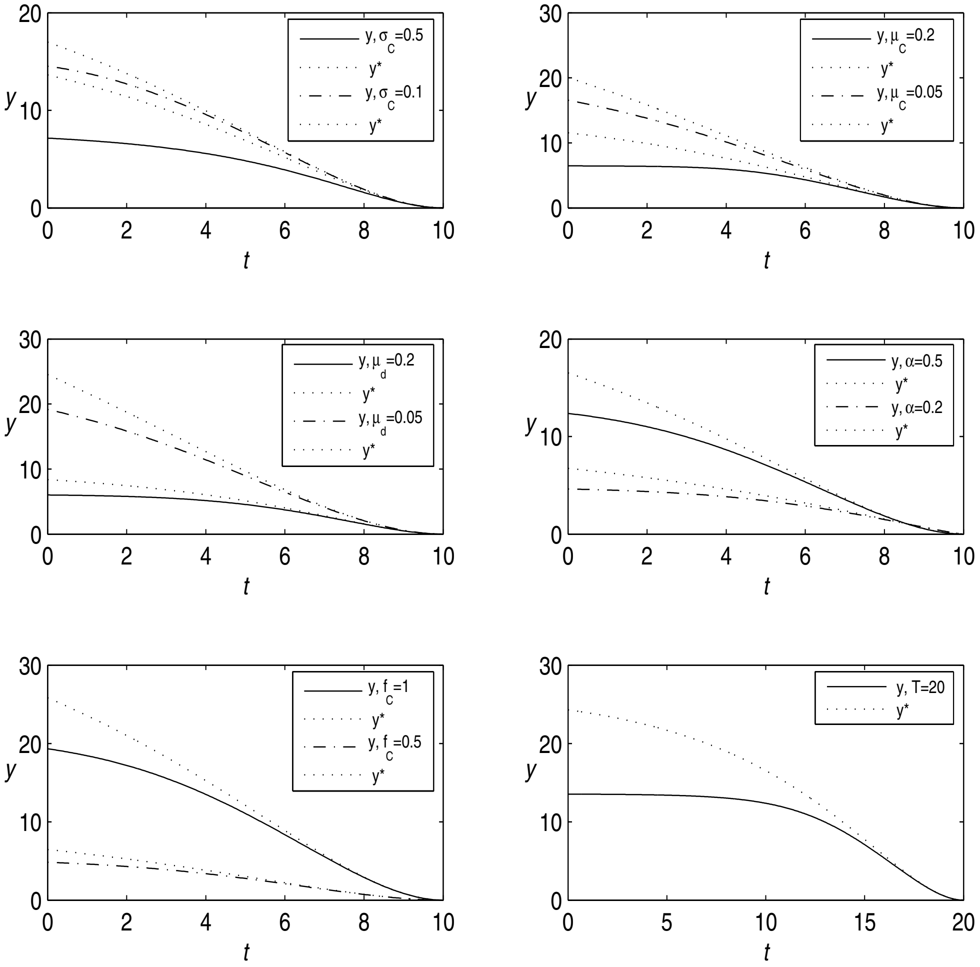


FIG. 1. $G \equiv 0$, $\sigma_C = 0.2$, $\mu_C = 0.1$, $\mu_d = 0.1$, $\alpha = 0.5$, $f_C = 0.8$, $T = 10$ *except as indicated in the legends.*

$$= \frac{\partial}{\partial C} u(x, C)\Big|_{x=\hat{x}(y,C)} + \vec{n}(\hat{x}(y, C))^\top \frac{\partial}{\partial C} \hat{x}(y, C),$$

where $\vec{n}(\hat{x}(y, C))$ is an outward normal to A at $\hat{x}(y, C)$; cf. [9, Corollary 3.9]. However, $\frac{\partial}{\partial C}\hat{x}(y, C) = \frac{\partial}{\partial C} I^{u_A(\cdot,C)}(y)$ is orthogonal to $\mathcal{N}_A(I^{u_A(\cdot,C)}(y))$ (see [9, Remark 3.15]), so we obtain $\vec{n}(\hat{x}(y, C))^\top \frac{\partial}{\partial C} I^{u_A(\cdot,C)}(y) = 0$, and (5.2) follows.

To prove strict concavity, we show that $\frac{\partial}{\partial C} u(x, C)|_{x=\hat{x}(y,C)}$ is strictly decreasing in $C \in \operatorname{int}(\mathcal{T})$. We write $H_u(x, C)$ for the Hessian of u with respect to *both variables* at (x, C), and $H_{u(\cdot,C)}(x)$ for the Hessian with respect to x only. For $y \in \mathcal{R}_{u_A}(C) := \nabla u(A \cap \operatorname{int}(\operatorname{dom}(u(\cdot, C))))$ fixed, (3.8) of [9] yields

$$\frac{d}{dC}\left(\frac{\partial}{\partial C} u(\hat{x}(y, C), C)\right) = \frac{\partial}{\partial C} \nabla u(x, C)^\top\Big|_{x=\hat{x}(y,C)} \hat{x}_C(y, C) + \frac{\partial^2}{\partial C^2} u(x, C)\Big|_{x=\hat{x}(y,C)}$$

$$(5.3) \qquad = -\frac{\partial}{\partial C} \nabla u(x, C)^\top\Big|_{x=\hat{x}(y,C)} \Big(H_{u(\cdot,C)}(\hat{x}(y, C))\Big)^{-1} \frac{\partial}{\partial C} \nabla u(x, C)\Big|_{x=\hat{x}(y,C)}$$

$$+ \frac{\partial^2}{\partial C^2} u(x,C)\Big|_{x=\hat{x}(y,C)}.$$

The previous expression is negative if $\frac{\partial}{\partial C}\nabla u = 0$ since u is strictly concave. Otherwise, let $\zeta = k^{-1}(H_{u(\cdot,C)})^{-1}\frac{\partial}{\partial C}\nabla u$ with $k = \sqrt{-\frac{\partial}{\partial C}\nabla u^\top (H_{u(\cdot,C)})^{-1}\frac{\partial}{\partial C}\nabla u}$. Since H_u is negative definite, then for any scalar s

$$\begin{aligned}
0 &> (s\,\zeta^\top, 1) H_u \begin{pmatrix} s\,\zeta \\ 1 \end{pmatrix} \\
&= -s^2 + 2s\,k^{-1}\frac{\partial}{\partial C}\nabla u(x,C)^\top\Big|_{x=\hat{x}(y,C)} \Big(H_{u(\cdot,C)}(\hat{x}(y,C))\Big)^{-1} \frac{\partial}{\partial C}\nabla u(x,C)\Big|_{x=\hat{x}(y,C)} \\
&\qquad + \frac{\partial^2}{\partial C^2} u(x,C)\Big|_{x=\hat{x}(y,C)} \\
&= -s^2 - 2s\,k + \frac{\partial^2}{\partial C^2} u(x,C)\Big|_{x=\hat{x}(y,C)}.
\end{aligned}$$

It follows that the discriminant of the above quadratic (in s) is negative, i.e.,

$$k^2 + \frac{\partial^2}{\partial C^2} u(x,C)\Big|_{x=\hat{x}(y,C)} < 0,$$

so the negativity of the last expression in (5.3) follows.

If $y \in \mathcal{S}_k(C)$, then the result follows from the same argument using (3.9) of [9] instead of (3.8) and

$$\zeta = k^{-1}\begin{pmatrix} ((H_{u(\cdot,C)})_{11})^{-1}\frac{\partial}{\partial C}\nabla_1 u \\ 0_{\dim(y^2)} \end{pmatrix}$$

for the appropriate $k = \sqrt{-\frac{\partial}{\partial C}\nabla_1 u^\top ((H_{u(\cdot,C)})_{11})^{-1}\frac{\partial}{\partial C}\nabla_1 u}$. Hence the proposition follows. □

It would appear that the results of [9, Proposition 3.14] and of Proposition 5.1 hold without the polyhedral property of A if the boundary of A is piecewise continuously differentiable.

REFERENCES

[1] A. B. Abel and J. C. Eberly, *An exact solution for the investment and value of a firm facing uncertainty, adjustment costs, and irreversibility*, J. Econ. Dynam. Control, 21 (1997), pp. 831–852.

[2] L. H. R. Alvarez, *Irreversible Investment, Incremental Capital Accumulation, and Price Uncertainty*, preprint, 2006.

[3] F. M. Baldursson and I. Karatzas, *Irreversible investment and industry equilibrium*, Finance Stoch., 1 (1997), pp. 69–89.

[4] G. Bertola, *Adjustment Costs and Dynamic Factor Demands: Investment and Employment under Uncertainty*, Ph.D. thesis, MIT, Cambridge, MA, 1988.

[5] G. Bertola, *Irreversible investments*, Research in Economics, 52 (1998), pp. 3–37.

[6] M. B. Chiarolla and U. G. Haussmann, *Explicit solution of a stochastic, irreversible investment problem and its moving threshold*, Math. Oper. Res., 30 (2005), pp. 91–108.

[7] M. B. Chiarolla and U. G. Haussmann, *Erratum*, Math. Oper. Res., 31 (2006), p. 432.

[8] M. B. Chiarolla and U. G. Haussmann, *A Stochastic Equilibrium Economy with Irreversible Investment*, preprint, 2008.

[9] M. B. Chiarolla and U. G. Haussmann, *Multivariable utility functions*, SIAM J. Optim., to appear.

[10] A. K. Dixit and R. S. Pindyck, *Investment under Uncertainty*, Princeton University Press, Princeton, NJ, 1994.

[11] X. Guo and H. Pham, *Optimal partially reversible investment with entry decision and general production functions*, Stochastic Processes Appl., 115 (2005), pp. 705–736.

[12] S. Hamadene and M. Jeanblanc, *On the starting and stopping problem: Application in reversible investments*, Math. Oper. Res., 32 (2007), pp. 182–192.

[13] S. Jacka, *Optimal stopping and the American put*, Math. Finance, 1 (1991), pp. 1–14.

[14] A. Merhi and M. Zervos, *A model for reversible investment capacity expansion*, SIAM J. Control Optim., 46 (2007), pp. 839–876.

[15] A. Oksendal, *Irreversible investment problems*, Finance Stoch., 4 (2000), pp. 223–250.

[16] B. Oksendal and K. Reikvam, *Viscosity solutions of optimal stopping problems*, Stochastics Stochastics Rep., 62 (1998), pp. 285–301.

[17] J. L. Pedersen and G. Peskir, *On nonlinear integral equations arising in problems of optimal stopping*, Proc. Funct. Anal. VII (Dubrovnik 2001), Various Publ. Ser. Aarhus, 46 (2002), pp. 159–175.

[18] F. Riedel and X. Su, *On Irreversible Investment*, preprint, 2006.

[19] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 2, pp. 463–480

© 2009 Society for Industrial and Applied Mathematics

REGULAR ALMOST INTERCONNECTION OF MULTIDIMENSIONAL BEHAVIORS*

ULRICH OBERST†

Abstract. Reading the doctoral thesis of Napp Avelli (2007) I realized that Gabriel's localization theory, which I applied in context with the stabilization of multidimensional input/output behaviors, can also be used for the construction of regular almost interconnections of behaviors in arbitrary dimensions and not only in two dimensions. In this paper I expose this theory in the language of quotient modules and derive an algorithm for arbitrary dimensions which has, however, not yet been implemented. Regular interconnections were introduced and discussed by Willems [*IEEE Trans. Automat. Control*, 36 (1991), pp. 259–294; 42 (1997), pp. 458–472] for one-dimensional behaviors. Their multidimensional counterparts have been treated by Rocha, Wood, Shankar, Zerz, Lomadze, Napp Avelli, and others since 1998. Two-dimensional almost direct sum decompositions and regular almost interconnections have been considered by Valcher and Bisiacco since 2000 and are also the main subject of Napp Avelli's thesis and his recent submitted paper. Roughly, two behaviors are almost equal if they differ by finite-dimensional behaviors only; so the latter are considered negligible in this context. Two-dimensional behaviors have special properties which are not discussed in the present paper. I also briefly discuss other variants of regular almost interconnections where only stable autonomous behaviors are considered negligible.

Key words. regular almost interconnection, multidimensional behavior, Gabriel localization

AMS subject classifications. 93B25, 93B40, 93C20

DOI. 10.1137/070707841

1. Introduction.

Regular interconnections were introduced and discussed by Willems in one dimension [23], [24], [16]. Soon thereafter Rocha and Wood [18], Shankar [20], and Zerz and Lomadze [26] extended the notions and partially the results to behaviors in higher dimensions. In context with the two-dimensional controllable-autonomous decomposition, almost direct sum decompositions have been discussed by Valcher and Bisiacco since 2000 [22], [1], [2]. Roughly, two behaviors are almost equal if they differ by finite-dimensional behaviors only; so the latter are considered negligible in this context. For such identifications the theory of quotient categories and modules was developed fifty years ago by Serre, Gabriel et al. That two-dimensional behaviors have various properties which do not hold in higher dimensions can be inferred from [4, Chap. VII, sec. 4, Thms. 2, 4, 5, and 6 and Ex. 3] and was also observed in [11, Thms. 7.42 and 7.74]. The starting point of the present paper was Chapter 5 of Napp Avelli's thesis [9], [10], where, in particular, the basic theory for arbitrary dimensions and an algorithm for two-dimensional regular almost interconnection are presented. Reading this thesis I realized that Gabriel localization of polynomial modules, which I applied in [14] for the stabilization of multidimensional input/output systems, can also be used to sharpen Napp Avelli's general theory and to extend his two-dimensional algorithm [10, Cor. 22] to arbitrary higher dimensions. In the present paper those results, for instance, [10, Thm. 17], which hold in two dimensions only are not discussed.

The regular almost interconnection problem is the following: Let $\mathcal{F}$ be one of the standard, discrete or continuous, multidimensional F-signal spaces over a field

[*] Received by the editors November 9, 2007; accepted for publication (in revised form) August 28, 2008; published electronically February 13, 2009.
http://www.siam.org/journals/sicon/48-2/70784.html

[†] Institut für Mathematik, Universität Innsbruck, Technikerstraße 25, A-6020 Innsbruck, Austria (Ulrich.Oberst@uibk.ac.at).

F and let $\mathcal{B} \subseteq \mathcal{B}_1, \mathcal{B}_2 \subseteq \mathcal{F}^l$ be behaviors defined as the solution spaces of linear systems of partial difference or differential equations with constant coefficients. Then $\mathcal{B}$ is called a *regular almost interconnection* of $\mathcal{B}_1$ and $\mathcal{B}_2$ if $\mathcal{F}^l = \mathcal{B}_1 + \mathcal{B}_2$ and if the behavior $(\mathcal{B}_1 \bigcap \mathcal{B}_2)/\mathcal{B}$ is finite-dimensional over F [10, Problem 19]. In the *almost* theory the behaviors $\mathcal{B}$ and $\mathcal{B}_1 \bigcap \mathcal{B}_2$ are then identified. The problem is to decide for given $\mathcal{B} \subseteq \mathcal{B}_1$ whether such a $\mathcal{B}_2$ exists and, if so, to construct one such $\mathcal{B}_2$. If in addition $\mathcal{B} = \mathcal{B}_1 \bigcap \mathcal{B}_2$, then $\mathcal{B}$ is called a regular interconnection of $\mathcal{B}_1$ and $\mathcal{B}_2$. The behavior $\mathcal{B}_1$ is interpreted as that of a given plant which by means of the controller $\mathcal{B}_2$ is changed to a desired behavior $\mathcal{B}$.

The main results of this paper are Theorem 3.3 and the associated Algorithm 4.1 for the solution of the regular almost interconnection problem in arbitrary dimensions. The latter makes essential use of Zerz and Lomadze's method [26, sec. 3] to check the existence of regular interconnections. The algorithmic problems in context with the stabilization of transfer matrices [7], [25] or input/output behaviors [14] do not appear here. Theorem 3.1 characterizes more general almost direct sum decompositions of behaviors [10, Problem 15] by means of quotient modules. But in contrast to the two-dimensional case of [10, Thm. 17] the corresponding algorithms in the general multidimensional situation of the present paper are not yet complete. Section 2 presents a survey without proofs of Gabriel's localization theory after [21] and [14, sec. 3] and derives several results which are needed for Theorems 3.1 and 3.3 and Algorithm 4.1. The short last section discusses other forms of regular almost interconnections where only stable autonomous behaviors are considered negligible.

2. Generalized quotient modules.

In the first part of this section we introduce the module categories which are relevant for the regular almost interconnection problem and give a survey of section 3 of [14], but refer to [21] and [14] for the details, the used (standard) terminology, and the notation. Corollary 2.1 describes the essential properties of the quotient module $Q(M)$ of a module M. In the second part we prove several results which are essential for the construction of a regular almost interconnection which itself is described in sections 3 and 4.

Let F be a field and $A := F[s] := F[s_1, \ldots, s_r]$ the polynomial algebra in $r \geq 2$ indeterminates with its quotient field $K := F(s)$. The category with the A-modules as objects and the A-linear maps as morphisms is denoted by Mod_A. The ring A is a factorial noetherian integral domain as assumed in section 3 of [14]. Let $\mathcal{F}$ be one of the *large injective cogenerator* signal A-modules from [11, main examples in sec. 2 and Thm. 2.54], for instance, the multisequence space $\mathcal{F} = \mathbb{C}^{\mathbb{N}^r}$, respectively, the space $\mathcal{F} = \mathcal{D}'(\mathbb{R}^r, \mathbb{C})$, of distributions in the discrete, respectively, continuous, cases of linear partial difference, respectively, differential, equations with constant coefficients. Its scalar multiplication is denoted by $\circ$. By definition an injective cogenerator is large if every finitely generated A-module M is a submodule of some finite power $\mathcal{F}^l$, up to isomorphism.

As in [14, sec. 3] let $\mathrm{Spec}(A)$, respectively, $\mathrm{Max}(A)$, denote the set of prime, respectively, of maximal, ideals of A. The *associator* or set of associated prime ideals of an A-module M is denoted by $\mathrm{Ass}(M)$. A prime ideal $\mathfrak{p}$ belongs to $\mathrm{Ass}(M)$ if $A/\mathfrak{p}$ is a submodule of M, up to isomorphism. The injective module $\mathcal{F}$ is a cogenerator if and only if it contains all simple modules $A/\mathfrak{m}$, $\mathfrak{m} \in \mathrm{Max}(A)$, up to isomorphism, i.e., if $\mathrm{Max}(A) \subseteq \mathrm{Ass}(\mathcal{F})$. It is a *large* injective cogenerator if and only if all modules $A/\mathfrak{p}$, $\mathfrak{p} \in \mathrm{Spec}(A)$, are contained in $\mathcal{F}$, up to isomorphism, or, in other terms, if $\mathrm{Ass}(\mathcal{F}) = \mathrm{Spec}(A)$ [11, Lem. 2.52]. In the complex continuous case the module

$$\mathcal{F}_{\mathrm{lf}} := \oplus_{\lambda \in \mathbb{C}^r} \mathbb{C}[t] e^{\lambda \bullet t}, \; t = (t_1, \ldots, t_r) \in \mathbb{R}^r, \; \lambda \bullet t := \lambda_1 t_1 + \cdots + \lambda_r t_r,$$

of *locally finite distributions* or *polynomial-exponential functions* is the unique least injective cogenerator [12, Thm. 6.6] and satisfies

$$\mathrm{Ass}(\mathcal{F}_{\mathrm{lf}}) = \mathrm{Max}(A),\ \text{ hence no } A/\mathfrak{p},\ \mathfrak{p} \in \mathrm{Spec}(A) \setminus \mathrm{Max}(A),$$

is contained in $\mathcal{F}_{\mathrm{lf}}$ and therefore $\mathcal{F}_{\mathrm{lf}}$ is definitely not large.

As the disjoint decomposition of $\mathrm{Spec}(A) = \mathrm{Ass}(\mathcal{F})$ according to [14, eq. (40)] we choose

$$\mathrm{Spec}(A) = \mathcal{P}_1 \uplus \mathcal{P}_2,\ \mathcal{P}_1 := \mathrm{Max}(A),\ \mathcal{P}_2 := \mathrm{Spec}(A) \setminus \mathrm{Max}(A). \tag{1}$$

Notice that any prime ideal $\mathfrak{p} = Ap$ of height one, where p is an irreducible polynomial, is not maximal in dimension $r \geq 2$ and therefore contained in $\mathcal{P}_2$. Since A is factorial we have

$$A = \bigcap_{p} \{A_{Ap};\ p \text{ irreducible}\} \subset K = F(s),\ \text{ hence also } A = \bigcap_{\mathfrak{p} \in \mathcal{P}_2} A_{\mathfrak{p}}. \tag{2}$$

A module M is called *locally finite* if its finitely generated submodules M' or, equivalently, its cyclic submodules Ax have finite F-dimension $[M' : F] := \dim_F(M')$. According to [4, Chap. IV, sec. 2.5, Prop. 7] or [13, Thm. 28] a module M is locally finite if and only if its associator $\mathrm{Ass}(M)$ or, equivalently, its support $\mathrm{supp}(M) := \{\mathfrak{p} \in \mathrm{Spec}(A);\ M_{\mathfrak{p}} \neq 0\}$ is contained in $\mathrm{Max}(A)$. The disjoint decomposition (1) gives rise to a direct sum decomposition

$$\begin{gathered} \mathcal{F} = \mathcal{F}_1 \oplus \mathcal{F}_2 \text{ with} \\ \mathrm{Ass}(\mathcal{F}_1) = \mathcal{P}_1 = \mathrm{Max}(A),\ \mathrm{Ass}(\mathcal{F}_2) = \mathcal{P}_2 = \mathrm{Spec}(A) \setminus \mathrm{Max}(A). \end{gathered} \tag{3}$$

Of course, both $\mathcal{F}_1$ and $\mathcal{F}_2$ are injective. For general decompositions $\mathrm{Spec}(A) = \mathcal{P}_1 \uplus \mathcal{P}_2$ as in [14, eq. (40)] the associated decomposition (3) is not unique. However, for the special decomposition of (1), $\mathcal{F}_1$ is uniquely determined and coincides with the locally finite part of $\mathcal{F}$. Indeed, according to [12, Thm. 1.14] we have

$$\begin{gathered} \mathcal{F}_1 = \mathcal{F}_{\mathrm{lf}} := \{y \in \mathcal{F};\ [A \circ y : F] < \infty\} = \bigoplus_{\mathfrak{m} \in \mathrm{Max}(A)} \mathcal{F}(\mathfrak{m}),\ \text{ where} \\ \mathcal{F}(\mathfrak{m}) = \bigcup_{k=0}^{\infty} \mathrm{ann}_{\mathcal{F}}(\mathfrak{m}^k),\ \mathrm{ann}_{\mathcal{F}}(\mathfrak{m}^k) := \{y \in \mathcal{F};\ \mathfrak{m}^k \circ y = 0\}. \end{gathered} \tag{4}$$

The module $\mathcal{F}(\mathfrak{m})$ is the indecomposable injective envelope of $A/\mathfrak{m}$ and even the least injective cogenerator over the local ring $A_{\mathfrak{m}}$. The injective module $\mathcal{F}_{\mathrm{lf}} = \mathcal{F}_1$ is the least injective cogenerator over $A = F[s]$. For many standard signal spaces $\mathcal{F}$ like $\mathbb{C}^{\mathbb{N}^r}$ or $\mathcal{D}'(\mathbb{R}^r, \mathbb{C})$ it coincides with the polynomial-exponential functions or sequences as derived in [12, Thms. 1.25, 5.26, 6.6, and 6.10]. While the direct complement $\mathcal{F}_2$ of $\mathcal{F}_{\mathrm{lf}}$ is not unique, the direct decomposition (3) implies the isomorphism of injective modules

$$\mathcal{F}_2 \cong \mathcal{F}/\mathcal{F}_{\mathrm{lf}}, \tag{5}$$

and therefore $\mathcal{F}_2$ is unique up to isomorphism. Notice, however, that in general no *constructive* description of a special $\mathcal{F}_2$ is available.

According to [14, eq. (50)] the decomposition (1) and the associated module $\mathcal{F}_2 \cong \mathcal{F}/\mathcal{F}_{\mathrm{lf}}$ give rise to a localization theory introduced by Gabriel [5]. More specifically we obtain the full *localizing or Serre subcategory* or *hereditary torsion class* $\mathfrak{C}$ of Mod_A and a *Gabriel topology* $\mathfrak{T}$ [21, Thm. VI.5.1] where

$$\begin{aligned}\mathfrak{C} &:= \{C \in \mathrm{Mod}_A;\ \mathrm{Hom}_A(C, \mathcal{F}_2) = 0\}\\ &= \{C \in \mathrm{Mod}_A;\ \forall \mathfrak{p} \in \mathcal{P}_2 :\ C_{\mathfrak{p}} = 0\},\\ \mathfrak{T} &:= \{\mathfrak{a} \subseteq A;\ A/\mathfrak{a} \in \mathfrak{C}\}\,.\end{aligned} \tag{6}$$

The modules in $\mathfrak{C}$ are called $\mathfrak{T}$-*torsion modules.* The second equality in (6) implies that

$$\mathfrak{C} := \{C;\ \mathrm{supp}(C) \subseteq \mathcal{P}_1 = \mathrm{Max}(A)\}. \tag{7}$$

This signifies that the $\mathfrak{T}$-torsion modules are exactly the locally finite ones. Either directly or by means of the injectivity of $\mathcal{F}_2$ one sees that the class $\mathfrak{C}$ is closed under taking submodules, factor modules, extensions, and direct sums, in particular

$$\mathfrak{C} = \{C \in \mathrm{Mod}_A;\ \forall x \in C :\ \mathrm{ann}_A(x) := \{a \in A; ax = 0\} \in \mathfrak{T}\}\,. \tag{8}$$

The largest submodule of M in $\mathfrak{C}$ or largest locally finite submodule M_{lf} of M is also called the $\mathfrak{T}$-*torsion radical* of M and denoted by $\mathrm{tor}_{\mathfrak{T}}(M) := M_{\mathrm{lf}}$. If $\mathrm{tor}_{\mathfrak{T}}(M) = 0$, the module is called $\mathfrak{T}$-*torsion free.* Due to (8) a $\mathfrak{T}$-torsion module is a torsion module in the usual sense, and hence a torsion free module is $\mathfrak{T}$-torsion free. The Chinese remainder theorem or [4, Chap. IV, sec. 2.5, Prop. 8] implies

$$M_{\mathrm{lf}} = \mathrm{tor}_{\mathfrak{T}}(M) = \oplus_{\mathfrak{m}\in \mathrm{Ass}(M)} M(\mathfrak{m}) \text{ with } M(\mathfrak{m}) := \bigcup_{k=0}^{\infty} \mathrm{ann}_M(\mathfrak{m}^k). \tag{9}$$

If ${}_AM$ is finitely generated, then $\mathrm{Ass}(M)$ is finite and the increasing sequence of annihilators $\mathrm{ann}_M(\mathfrak{m}^k)$ becomes stationary. If $e(\mathfrak{m})$ is the least index k such that $\mathrm{ann}_M(\mathfrak{m}^k) = \mathrm{ann}_M(\mathfrak{m}^{k+1})$, we obtain

$$\begin{gathered} M_{\mathrm{lf}} = \mathrm{tor}_{\mathfrak{T}}(M) = \oplus_{\mathfrak{m}\in \mathrm{Ass}(M)} M(\mathfrak{m}) \text{ and}\\ 0 = \mathrm{ann}_M(\mathfrak{m}^0) \subsetneq \mathrm{ann}_M(\mathfrak{m}^1) \subsetneq \cdots \subsetneq \mathrm{ann}_M(\mathfrak{m}^{e(\mathfrak{m})}) = M(\mathfrak{m}),\ e(\mathfrak{m}) > 0. \end{gathered} \tag{10}$$

An A-module N is called $\mathfrak{T}$-*closed* if for every ideal $\mathfrak{a} \in \mathfrak{T}$ the canonical map

$$\mathrm{Hom}(\mathrm{inj}, N) : N = \mathrm{Hom}_A(A, N) \to \mathrm{Hom}_A(\mathfrak{a}, N),\ x \mapsto (a \mapsto ax) \tag{11}$$

is an isomorphism or, in other words, if for every linear map $f : \mathfrak{a} \to N$ there is a unique $x \in N$ with $f(a) = ax$ for all $a \in \mathfrak{a}$. The full additive subcategory of Mod_A of all $\mathfrak{T}$-closed submodules is denoted by $\mathrm{Mod}_{A,\mathfrak{T}}$. Its properties are described in [21, Chap. X, sec. 1]. It is obviously closed under taking arbitrary inverse limits or, equivalently, under products and kernels and therefore the kernels, products, and limits in $\mathrm{Mod}_{A,\mathfrak{T}}$ coincide with the standard ones in Mod_A. The category $\mathrm{Mod}_{A,\mathfrak{T}}$ also admits arbitrary colimits or, equivalently, coproducts or direct sums and cokernels and is indeed abelian. According to [21, pp. 195–200, 213–216] the inclusion functor $\mathrm{inj} : \mathrm{Mod}_{A,\mathfrak{T}} \subset \mathrm{Mod}_A$ has an exact left adjoint *quotient module or localization* functor

$$Q : \mathrm{Mod}_A \to \mathrm{Mod}_{A,\mathfrak{T}},\ M \mapsto Q(M), \tag{12}$$

with the functorial adjunction morphism

$$\eta_M : M \to Q(M), \text{ i.e., } \mathrm{Hom}_A(Q(M), N) \cong \mathrm{Hom}_A(M, N),\ g \mapsto g\eta_M, \tag{13}$$

for $M \in \mathrm{Mod}_A$ and $N \in \mathrm{Mod}_{A,\mathfrak{T}}$. In [14, sec. 3] we used the notation $M_{\mathfrak{T}} := Q(M)$. The functor Q is unique up to functorial isomorphism. Concrete representations of $Q(M)$ in various cases will be given below.

The inclusion functor $\mathrm{inj} : \mathrm{Mod}_{A,\mathfrak{T}} \to \mathrm{Mod}_A$ preserves all limits as a right adjoint functor and is left exact in particular, but not right exact. If N_1 and N_2 are $\mathfrak{T}$-closed and $g : N_1 \to N_2$ is A-linear, the cokernel of g in the category $\mathrm{Mod}_{A,\mathfrak{T}}$ is given as

$$\mathrm{cok}_{\mathfrak{T}}(g : N_1 \to N_2) = Q\left(N_2/g(N_1)\right). \tag{14}$$

The kernel of the canonical map $\eta_M : M \to Q(M)$ is [21, Chap. IX, Lem. 1.2]

$$M_{\mathrm{lf}} = \mathrm{tor}_{\mathfrak{T}}(M) = \ker(\eta_M : M \to Q(M)). \tag{15}$$

Hence η_M is a monomorphism and then $M \subset Q(M)$ by identification if and only if M is $\mathfrak{T}$-torsionfree. The adjointness of Q and the properties of $\mathfrak{C}$ directly imply

$$\begin{gathered} Q(N) = N \text{ for all } N \in \mathrm{Mod}_{A,\mathfrak{T}} \text{ and} \\ Q(N) = 0 \iff N \in \mathfrak{C} \iff N \text{ is locally finite, hence} \\ Q(M) = Q(M/\mathrm{tor}_{\mathfrak{T}}(M)) \text{ for all } M \in \mathrm{Mod}_A\,. \end{gathered} \tag{16}$$

In particular, a $\mathfrak{T}$-closed module is $\mathfrak{T}$-torsionfree.

An arbitrary morphism $f : M_1 \to M_2$ in Mod_A gives rise to the exact sequences

$$\begin{gathered} 0 \to \ker(f) \xrightarrow{\mathrm{inj}} M_1 \xrightarrow{f} M_2 \xrightarrow{\mathrm{can}} \mathrm{cok}(f) = M_2/f(M_1) \to 0 \text{ in } \mathrm{Mod}_A \text{ and} \\ 0 \to Q(\ker(f)) \xrightarrow{\mathrm{inj}} Q(M_1) \xrightarrow{Q(f)} Q(M_2) \xrightarrow{Q(\mathrm{can})} Q(\mathrm{cok}(f)) \to 0 \text{ in } \mathrm{Mod}_{A,\mathfrak{T}}\,. \end{gathered} \tag{17}$$

Hence

$$Q(f) \text{ is } \begin{cases} \text{zero} \\ \text{a monomorphism} \\ \text{an epimorphism} \\ \text{an isomorphism} \end{cases} \iff \begin{cases} \mathrm{im}(f) \in \mathfrak{C} \\ \ker(f) \in \mathfrak{C} \\ \mathrm{cok}(f) \in \mathfrak{C} \\ \ker(f),\ \mathrm{cok}(f) \in \mathfrak{C}. \end{cases} \tag{18}$$

If $\mathrm{im}(f) \in \mathfrak{C}$, the map f is called $\mathfrak{T}$-*zero* or *almost zero* or *zero modulo* $\mathfrak{C}$. The *almost* terminology is due to Napp Avelli [9], [10]. Analogously we define $\mathfrak{T}$- or almost monomorphisms, epimorphisms, and isomorphisms. In dimension $r = 2$ these maps coincide with the *pseudozero, pseudoinjective maps*, etc., from [4, Chap. VII, sec. 4.4, Def. 3]. In particular,

$$\eta_M : M \to Q(M) \text{ with } Q(\eta_M) = \mathrm{id}_{Q(M)} : Q(M) \to Q(Q(M)) = Q(M) \tag{19}$$

is a $\mathfrak{T}$-isomorphism.

Corollary 2.1. *The quotient functor Q is characterized by the the following equivalent properties, i.e., every additive functor $Q_1 : \mathrm{Mod}_A \to \mathrm{Mod}_{A,\mathfrak{T}}$ with these properties coincides with Q up to a functorial isomorphism:*

1. *The adjointness relation* (13) *holds for Q_1.*

2. $Q_1 : \mathrm{Mod}_A \to \mathrm{Mod}_{A,\mathfrak{T}}$ *is exact and* (16) *holds for* Q_1.
3. $Q_1(N) = N$ *for each* $\mathfrak{T}$*-closed module* N*, and* $Q_1(f)$ *is an isomorphism if* f *is a* $\mathfrak{T}$*-isomorphism.*

In particular, a module M *is annihilated by* Q*, i.e.,* $Q(M) = 0$*, if and only if it is locally finite. Thus application of* Q *signifies to ignore locally finite and especially finite-dimensional* A*-modules. Most of the subsequent derivations except* (20) *and* (21) *use the preceding properties only.*

According to [14, Lems. 3.2 and 3.4] and (2) the quotient module $Q(A)$ of A is

$$Q(A) = \bigcap_{\mathfrak{p}\in\mathcal{P}_2} A_{\mathfrak{p}} = A \subset K = F(s) \tag{20}$$

and thus coincides with A. If, more generally, U is a finitely generated torsionfree module and thus a submodule of some $A^{1\times m}$, the quotient $Q(U)$ is given [14, Lem. 3.6] as

$$U \subseteq Q(U) = \bigcap_{\mathfrak{p}\in\mathcal{P}_2} U_{\mathfrak{p}} \subseteq Q(A^{1\times m}) = A^{1\times m}. \tag{21}$$

In particular, $Q(U)$ is itself a finitely generated A-module. The canonical Gel'fand map

$$U \to U^{**} := \mathrm{Hom}_A(\mathrm{Hom}_A(U, A), A),\ u \mapsto (\alpha \mapsto \alpha(u)),$$

is injective and U^{**} is naturally identified with its bidual lattice in $A^{1\times m}$ [4, Chap. VII, p. 517]. Then [4, Chap. VII, sec. 4.2, Thm. 2]

$$U \subseteq Q(U) = \bigcap_{\mathfrak{p}\in\mathcal{P}_2} U_{\mathfrak{p}} \subseteq \bigcap_{p \text{ irreducible}} U_{Ap} = U^{**} \subseteq A^{1\times m}. \tag{22}$$

Therefore, if U is reflexive, i.e., if the Gel'fand map is an isomorphism or $U = U^{**}$, we get that $Q(U) = U$ and U is $\mathfrak{T}$-closed. In particular, the dual module $M^* = (M/\operatorname{tor}(M))^*$ and the bidual module M^{**} are reflexive and thus $\mathfrak{T}$-closed for any finitely generated A-module M. If M is a finitely generated torsion module, the quotient $Q(M) = Q(M/M_{lf})$ is not finitely generated in general, and this creates problems for explicit computations.

Remark 2.2. In dimension $r = 2$ we have

$$\mathcal{P}_2 = \mathrm{Spec}(A) \setminus \mathrm{Max}(A) = \{Ap;\ p \text{ irreducible}\} \cup \{0\},$$

and therefore $Q(U) = U^{**}$ for a finitely generated torsionfree U. From the general theory we conclude that $U^{**}/U = Q(U)/U$ is finite-dimensional. The module $U^+ := U^{**}$ is free and plays an important part in Napp Avelli's two-dimensional theory [10, p. 7, Thm. 12, Lem. 21].

The module $\mathcal{F}_2 \cong \mathcal{F}/\mathcal{F}_{\mathrm{lf}}$ is an injective cogenerator of $\mathrm{Mod}_{A,\mathfrak{T}}$ [21, Chap. X, Prop. 1.9].

Lemma 2.3 (cf. [21, Chap. IX, Prop. 2.1]). *Let* N *be a* $\mathfrak{T}$*-closed module and* $f : M_1 \to M_2$ *a* $\mathfrak{T}$*-isomorphism, i.e., a linear map with* $\ker(f)$, $\mathrm{cok}(f) \in \mathfrak{C}$.

1. *The canonical map*

$$\mathrm{Hom}(f, N) : \mathrm{Hom}_A(M_2, N) \to \mathrm{Hom}_A(M_1, N),\ g_2 \mapsto g_1 := g_2 f,$$

is an isomorphism or, equivalently, any linear map $g_1 : M_1 \to N$ *can be uniquely extended to* $g_2 : M_2 \to N$ *with* $g_2 f = g_1$.

2. *If $M_1 \subseteq M_2$, $M_2/M_1 \in \mathfrak{C}$, and $\operatorname{tor}_{\mathfrak{T}}(M_2) = 0$, the inclusion $M_1 \subseteq M_2$ is essential; i.e., for each nonzero submodule U of M_2 also $U \bigcap M_1$ is nonzero. In particular, the inclusion $M_2 \subseteq Q(M_2)$ is essential.*

Proof.

1. The adjointness isomorphism $\operatorname{Hom}_A(Q(M), N) \cong \operatorname{Hom}_A(M, N)$ implies

$$\operatorname{Hom}_A(M_2, N) \cong \operatorname{Hom}_A(Q(M_2), N) \overset{\operatorname{Hom}(Q(f),N)}{\cong},$$
$$\operatorname{Hom}_A(Q(M_1), N) \cong \operatorname{Hom}_A(M_1, N).$$

2. Let $0 \neq x \in U$. Since $\overline{x} \in M_2/M_1 \in \mathfrak{C}$ the annihilator $\mathfrak{a} := \operatorname{ann}_A(\overline{x})$ belongs to $\mathfrak{T}$ and $\mathfrak{a}x \subseteq M_1$. The condition

$$\operatorname{tor}_{\mathfrak{T}}(M_2) = 0 \quad \text{implies} \quad 0 \neq \mathfrak{a}x \subseteq U \bigcap M_1$$

and thus the assertion. □

Theorem 2.4. *If M is a submodule of the $\mathfrak{T}$-closed module N, then*

$$M \subseteq Q(M) \subseteq N \text{ and } Q(M)/M = \operatorname{tor}_{\mathfrak{T}}(N/M).$$

In other words, $Q(M)$ is the largest submodule V with $M \subseteq V \subseteq N$ and locally finite V/M, and hence

$$\begin{aligned} Q(M) &= \{y \in N;\ \exists \mathfrak{a} \in \mathfrak{T} \text{ with } \mathfrak{a}y \subseteq M\} \\ &= \{y \in N;\ \dim_F((Ay + M)/M) < \infty\}. \end{aligned}$$

Notice here that Q is unique up to a functorial isomorphism only. But as submodule of N the quotient module $Q(M)$ is uniquely determined as the largest submodule with locally finite factor module.

Proof.

1. Since N is $\mathfrak{T}$-closed its $\mathfrak{T}$-torsion submodule is zero, and the same holds for M. We infer $M \subseteq Q(M)$ and $Q(M)/M \in \mathfrak{C}$. By item 1 of Lemma 2.3 the injection $\operatorname{inj} : M \to N$ has a unique extension

$$g : Q(M) \to N \text{ with } \ker(g) \bigcap M = \ker(\operatorname{inj}) = 0, \text{ hence } \ker(g) = 0$$

according to item 2 of the lemma. Thus g is a monomorphism and

$$Q(M) \cong g(Q(M)) \text{ and } M = g(M) \subseteq g(Q(M)) \subseteq N,\ g(Q(M))/M \in \mathfrak{C}.$$

Without loss of generality we therefore assume

$$M \subseteq Q(M) \subseteq N \text{ and } Q(M)/M \subseteq \operatorname{tor}_{\mathfrak{T}}(N/M).$$

2. Let $U \supseteq M$ be the unique submodule of N with $U/M = \operatorname{tor}_{\mathfrak{T}}(N/M)$, and hence $Q(M) \subseteq U$. Since U/M belongs to $\mathfrak{C}$ so does $U/Q(M)$. Since $Q(M)$ is $\mathfrak{T}$-closed the identity map $\operatorname{id}_{Q(M)}$ has a unique extension $h : U \to Q(M)$, again by Lemma 2.3, item 1, and then

$$U = Q(M) \oplus \ker(h) \text{ and } \ker(h) \bigcap Q(M) = 0.$$

Since the inclusion $Q(M) \subseteq U$ is essential according to item 2 of the lemma we infer $\ker(h) = 0$ and $U = Q(M)$ as asserted. □

THEOREM 2.5. *Let M be a finitely generated A-module with*

$$M_{\mathrm{lf}} = \mathrm{tor}_{\mathfrak{T}}(M) = 0 \text{ or } \mathrm{Ass}(M) \subseteq \mathcal{P}_2 = \mathrm{Spec}(A) \setminus \mathrm{Max}(A).$$

1. *The set $S := \bigcap_{\mathfrak{p}\in\mathrm{Ass}(M)}(A \setminus \mathfrak{p}) \subset A$ is multiplicatively closed and the quotient ring*

$$A_S = \bigcap_{\mathfrak{p}\in\mathrm{Ass}(M)} A_{\mathfrak{p}} \subseteq K = F(s) \text{ is semilocal with}$$
$$\mathrm{Max}(A_S) = \{\mathfrak{p}_S;\ \mathfrak{p} \text{ is maximal in } \mathrm{Ass}(M)\}.$$

2. *The canonical map $M \to M_S$ is injective.*
3. *The module M_S is $\mathfrak{T}$-closed as A-module.*

By Theorem 2.4 this implies

$$M \subseteq Q(M) \subseteq M_S,\ Q(M)/M = \mathrm{tor}_{\mathfrak{T}}(M_S/M) \text{ and}$$
$$Q(M) = \{y \in M_S;\ \dim_F\left((Ay + M)/M\right) < \infty\}. \tag{23}$$

Proof.
1. The first assertion follows from [4, Chap. II, sec. 3.5, Prop. 17].
2. Corollary 2 from [4, Chap. IV, sec. 1.2] implies

$$S := \bigcap_{\mathfrak{p}\in\mathrm{Ass}(M)} (A \setminus \mathfrak{p}) = \{s \in A;\ s\circ : M \to M \text{ is injective}\} \text{ and thus also}$$
$$\ker\left(\mathrm{can} : M \to M_S,\ x \mapsto \frac{x}{1}\right) = \{x \in M;\ \exists s \in S \text{ with } sx = 0\} = 0.$$

Therefore we can and do identify $M \subseteq M_S,\ x = \frac{x}{1}$.

3. Let $\mathfrak{a} \in \mathfrak{T}$ and thus

$$\mathrm{supp}(A/\mathfrak{a}) = \{\mathfrak{p} \in \mathrm{Spec}(A);\ \mathfrak{a} \subseteq \mathfrak{p}\} \subseteq \mathrm{Max}(A).$$

We show that $\mathfrak{a} \cap S \neq \emptyset$ and therefore $\mathfrak{a}_S = A_S$. If

$$\mathfrak{a} \cap S = \emptyset, \text{ then } \mathfrak{a} \subseteq \bigcup_{\mathfrak{p}\in\mathrm{Ass}(M)} \mathfrak{p} \text{ and } \mathfrak{a} \subseteq \mathfrak{p}_0$$

for some $\mathfrak{p}_0 \in \mathrm{Ass}(M)$, the last implication following from [4, Chap. II, sec. 1.2, Prop. 2]. Since $\mathfrak{a} \in \mathfrak{T}$ this prime ideal $\mathfrak{p}_0$ is maximal, and since $\mathrm{Ass}(M) \subset \mathrm{Spec}(A) \setminus \mathrm{Max}(A)$ it is not. This is a contradiction and therefore $\mathfrak{a}_S = A_S$. But then

$$M_S \cong \mathrm{Hom}_{A_S}(A_S, M_S) = \mathrm{Hom}_{A_S}(\mathfrak{a}_S, M_S) \cong \mathrm{Hom}_A(\mathfrak{a}, M_S) \text{ for all } \mathfrak{a} \in \mathfrak{T}.$$

This signifies that M_S is $\mathfrak{T}$-closed. □

3. Regular almost interconnections. The assumptions are the same as in the preceding section. The next theorem characterizes the almost direct sum decomposition of modules and behaviors [10, Problem 15, Thm. 17] by means of quotient modules, but does *not* contain the two-dimensional constructive part of [10, Thm. 17].

THEOREM 3.1. *For a finitely generated A-module M and a submodule M_1 the following assertions are equivalent:*

1. *There is a submodule M_2 of M such that the canonical map*

$$+ : M_1 \times M_2 \xrightarrow{+} M,\ (x_1, x_2) \mapsto x_1 + x_2,\ \textit{is a } \mathfrak{T}\textit{-isomorphism},$$

i.e., has F-finite dimensional kernel, isomorphic to $M_1 \bigcap M_2$, and cokernel $M/(M_1 + M_2)$.

2. *$Q(M_1)$ is a direct summand of $Q(M) = Q(M/\operatorname{tor}_{\mathfrak{T}}(M))$.*

If these conditions are satisfied and if M is $\mathfrak{T}$-torsionfree, i.e., $M_{\mathrm{lf}} = \operatorname{tor}_{\mathfrak{T}}(M) = 0$, then the map $+$ is injective, i.e., $M_1 \bigcap M_2 = 0$ and $M_1 + M_2 = M_1 \oplus M_2$.

Proof. $1 \implies 2$: Recall that Q is exact and that thus $Q(M_1)$ is a submodule of $Q(M)$. The $\mathfrak{T}$-isomorphism implies the isomorphism

$$+ = Q(+) : Q(M_1 \times M_2) = Q(M_1) \times Q(M_2) \cong Q(M), \text{ and hence}$$
$$Q(M_1) \oplus Q(M_2) = Q(M).$$

$2 \implies 1$: Assume $Q(M_1) \oplus V = Q(M)$. As a direct summand of a $\mathfrak{T}$-closed module V is also $\mathfrak{T}$-closed. Let $g := \eta_M : M \to Q(M)$ be the universal map which is a $\mathfrak{T}$-isomorphism and hence has locally finite kernel and cokernel. Define $M_2 := g^{-1}(V) \subseteq M$ and the restriction $g|M_2 : M_2 \to V$. With g also its restriction has locally finite kernel and cokernel, is therefore a $\mathfrak{T}$-isomorphism too, and induces the isomorphism $Q(g|M_2) : Q(M_2) \cong Q(V) = V$. Summing up we obtain commutative diagrams

$$\begin{tikzcd} M_1 \times M_2 \arrow[r, "+"] \arrow[d, "\eta_{M_1} \times g|M_2"'] & M \arrow[d, "g"] \\ Q(M_1) \times V \arrow[r, "+"] & Q(M) \end{tikzcd} \quad \text{and} \quad \begin{tikzcd} Q(M_1) \times Q(M_2) \arrow[r, "+"] \arrow[d, "Q(\eta_{M_1}) \times Q(g|M_2)"'] & Q(M) \arrow[d, "Q(g)"] \\ Q(M_1) \times V \arrow[r, "+"] & Q(M) \end{tikzcd},$$

where the right diagram is obtained from the left one by application of Q, where $Q(N) = N$ for each $\mathfrak{T}$-closed module, and where the vertical maps and the lower horizontal map in the right diagram are bijective by construction, respectively, due to $Q(M_1) \oplus V = Q(M)$. Therefore all maps in the right diagram are isomorphisms and hence all morphisms in the left diagram and especially $+ : M_1 \times M_2 \to M$ are $\mathfrak{T}$-isomorphisms as asserted.

By construction $M_1 \bigcap M_2 \cong \ker(+)$ is F-finite-dimensional and thus a $\mathfrak{T}$-torsion submodule of M whose largest $\mathfrak{T}$-torsion submodule is $\operatorname{tor}_{\mathfrak{T}}(M)$ and zero by assumption. This implies $M_1 \bigcap M_2 = 0$. □

The behavioral interpretation of the preceding theorem is the following: Let

$$M = A^{1\times l}/U \supset M_i = U_i/U,\ i = 1, 2, \text{ and hence}$$

$$\ker(M_1 \times M_2 \xrightarrow{+} M) \cong M_1 \bigcap M_2 = \left(U_1 \bigcap U_2\right)/U, \tag{24}$$

$$\operatorname{cok}(M_1 \times M_2 \xrightarrow{+} M) = M/(M_1 + M_2) \cong A^{1\times l}/(U_1 + U_2).$$

The modules in the last two rows are F-finite-dimensional by Theorem 3.1. Let $D := \operatorname{Hom}_A(-, \mathcal{F})$ denote the duality functor. For the standard injective cogenerators from [11] the module M is finite-dimensional if and only if $D(M)$ has this property [13, Thm. 17], and then

$$\dim_F(M) = \dim_F(D(M)). \tag{25}$$

The preceding modules give rise to the behaviors [11, Cor. 2.48]

$$\begin{gathered}\mathcal{B} := U^{\perp} := \{w \in \mathcal{F}^l;\ U \circ w = 0\} \cong D(M) = \mathrm{Hom}_A(A^{1\times l}/U, \mathcal{F}) \text{ with}\\ U = \mathcal{B}^{\perp} := \{\xi \in A^{1\times l};\ \xi \circ \mathcal{B} = 0\},\\ \mathcal{B}_i := U_i^{\perp} \subseteq \mathcal{B} = U^{\perp},\ \mathcal{B}_1 \bigcap \mathcal{B}_2 = (U_1 + U_2)^{\perp},\ \mathcal{B}_1 + \mathcal{B}_2 = \left(U_1 \bigcap U_2\right)^{\perp},\\ \mathcal{B}/(\mathcal{B}_1 + \mathcal{B}_2) \cong D\left(\left(U_1 \bigcap U_2\right)/U\right),\ \mathcal{B}_1 \bigcap \mathcal{B}_2 \cong D\left(A^{1\times l}/(U_1 + U_2)\right).\end{gathered} \tag{26}$$

Comparison of (24) and (26) by means of (25) furnishes the following.

Corollary and Definition 3.2 ($\mathfrak{T}$- or almost direct sum decomposition). *For behaviors $\mathcal{B}_1, \mathcal{B}_2 \subseteq \mathcal{B} \subseteq \mathcal{F}^l$ the behaviors $\mathcal{B}_1 \bigcap \mathcal{B}_2$ and $\mathcal{B}/(\mathcal{B}_1 + \mathcal{B}_2)$ are finite-dimensional if and only if this holds for the modules*

$$\begin{gathered}A^{1\times l}/\left(\mathcal{B}_1^{\perp} + \mathcal{B}_2^{\perp}\right) \text{ and } \left(\mathcal{B}_1^{\perp} \bigcap \mathcal{B}_2^{\perp}\right)/\mathcal{B}^{\perp}, \text{ and then}\\ \dim_F\left(\mathcal{B}_1 \bigcap \mathcal{B}_2\right) = \dim_F\left(A^{1\times l}/\left(\mathcal{B}_1^{\perp} + \mathcal{B}_2^{\perp}\right)\right) \text{ and}\\ \dim_F\left(\mathcal{B}/\left(\mathcal{B}_1 + \mathcal{B}_2\right)\right) = \dim_F\left(\left(\mathcal{B}_1^{\perp} \bigcap \mathcal{B}_2^{\perp}\right)/\mathcal{B}^{\perp}\right).\end{gathered}$$

Under these equivalent conditions $\mathcal{B}$ is called the $\mathfrak{T}$- or almost direct sum *of the two subbehaviors $\mathcal{B}_1$ and $\mathcal{B}_2$* [10, Problem 15]. *For $r = 2$ Theorem* 17 *of* [10] *characterizes these decompositions constructively.*

If $A^{1\times l}/\mathcal{B}^{\perp}$ is $\mathfrak{T}$-torsionfree, i.e., if $\mathcal{B}$ has no finite-dimensional factor behavior, then the equality $\mathcal{B} = \mathcal{B}_1 + \mathcal{B}_2$ holds. The paper [2] *discusses almost direct sum decompositions for $r = 2$ with this additional property $\mathcal{B}_1 + \mathcal{B}_2 = \mathcal{B}$.*

For the regular almost interconnection problem we assume that

$$\begin{gathered}\mathcal{B} \subseteq \mathcal{B}_1,\ \mathcal{B}_2 \subseteq \mathcal{F}^l,\ \mathcal{B} := U^{\perp},\ \mathcal{B}_i := U_i^{\perp}, \text{ and hence } U = \mathcal{B}^{\perp},\ U_i = \mathcal{B}_i^{\perp},\\ \mathcal{B}_1 \bigcap \mathcal{B}_2 = (U_1 + U_2)^{\perp},\ \left(\mathcal{B}_1 \bigcap \mathcal{B}_2\right)/\mathcal{B} \cong D\left(U/(U_1 + U_2)\right),\\ \mathcal{B}_1 + \mathcal{B}_2 = \left(U_1 \bigcap U_2\right)^{\perp} \cong D\left(A^{1\times l}/\left(U_1 \bigcap U_2\right)\right) \subseteq \mathcal{F}^l = D(A^{1\times l}).\end{gathered} \tag{27}$$

Theorem 3.3 (regular $\mathfrak{T}$- or almost interconnection). *For subbehaviors $\mathcal{B} \subseteq \mathcal{B}_1 \subseteq \mathcal{F}^l$ and the just introduced notations, the following assertions are equivalent:*

1. *There is a submodule U_2 of U such that the canonical map $U_1 \times U_2 \xrightarrow{+} U$ is a $\mathfrak{T}$-isomorphism, and then $U_1 \bigcap U_2 = 0$.*
2. *$Q(U_1)$ is a direct summand of $Q(U) \subseteq A^{1\times l}$.*
3. *There is a subbehavior $\mathcal{B}_2$ of $\mathcal{F}^l$ which also contains $\mathcal{B}$ such that $\mathcal{B}_1 + \mathcal{B}_2 = \mathcal{F}^l$ and $(\mathcal{B}_1 \bigcap \mathcal{B}_2)/\mathcal{B}$ is finite-dimensional.*
4. *There is a subbehavior $\mathcal{B}_2$ of $\mathcal{F}^l$ which also contains $\mathcal{B}$ such that*

$$\mathcal{B}_1 + \mathcal{B}_2 = \mathcal{F}^l \text{ and } \mathcal{B}_1 \bigcap \mathcal{B}_2 \bigcap \mathcal{F}_2^l = \mathcal{B} \bigcap \mathcal{F}_2^l.$$

If the preceding equivalent conditions are satisfied, the behavior $\mathcal{B}$ is called a regular $\mathfrak{T}$- or almost interconnection *of $\mathcal{B}_1$ and $\mathcal{B}_2$. Item* 4 *is of limited value since a direct summand $\mathcal{F}_2$ cannot be determined constructively.*

In dimension $r = 2$ we have $Q(U_1) = U_1^{+}$ and $Q(U) = U^{+}$ according to Remark 2.2. *Napp Avelli's criterion* [10, Lem. 21] *thus coincides with item* 2.

Proof. 1 $\iff$ 2: Theorem 3.1. Moreover, since U is torsionfree and $U_1 \bigcap U_2 \cong \ker(+)$ is finite-dimensional the identity $U_1 \bigcap U_2 = 0$ follows.

2 $\iff$ 3: This follows like Corollary 3.2 from (27) and the equivalence

$$U_1 \bigcap U_2 = 0 \iff \mathcal{B}_1 + \mathcal{B}_2 = \left(U_1 \bigcap U_2\right)^{\perp} = 0^{\perp} = \mathcal{F}^l.$$

3 $\implies$ 4: The decomposition $\mathcal{F} = \mathcal{F}_{\mathrm{lf}} \oplus \mathcal{F}_2$ from (3) induces decompositions

$$\mathcal{B} \cong \mathrm{Hom}_A(A^{1\times l}/U, \mathcal{F}) \cong \mathrm{Hom}_A(A^{1\times l}/U, \mathcal{F}_{\mathrm{lf}}) \oplus \mathrm{Hom}_A(A^{1\times l}/U, \mathcal{F}_2),$$

$$\text{and hence } \mathcal{B} = \left(\mathcal{B} \bigcap \mathcal{F}_{\mathrm{lf}}^l\right) \oplus \left(\mathcal{B} \bigcap \mathcal{F}_2^l\right) \text{ with } \mathcal{B} \bigcap \mathcal{F}_2^l \cong \mathrm{Hom}_A(A^{1\times l}/U, \mathcal{F}_2)$$

and analogous decompositions for all other behaviors. The $\mathfrak{T}$-torsion modules C are characterized by $\mathrm{Hom}_A(C, \mathcal{F}_2) = 0$ according to (6), in particular $\mathrm{Hom}_A(U/(U_1 + U_2), \mathcal{F}_2) = 0$. Since $\mathcal{F}_2$ is injective the functor $D_2 := \mathrm{Hom}_A(-, \mathcal{F}_2)$ is exact. Therefore the canonical exact sequence

$$0 \to U/(U_1 + U_2) \to A^{1\times l}/(U_1 + U_2) \to A^{1\times l}/U \to 0$$

induces the exact sequence

$$0 \to D_2(A^{1\times l}/U) \to D_2(A^{1\times l}/(U_1 + U_2)) \to D_2(U/(U_1 + U_2) = 0,$$

$$\text{and hence } \mathcal{B} \bigcap \mathcal{F}_2^l = \mathcal{B}_1 \bigcap \mathcal{B}_2 \bigcap \mathcal{F}_2^l.$$

4 $\implies$ 3: analogous. □

4. Algorithms. In this section we prove Algorithm 4.1, which makes Theorem 3.3 constructive, and thereby extend the algorithm of Napp Avelli [10, Cor. 22] to higher dimensions than two, but with a completely different method. The algorithm makes essential use of the algorithm of Zerz and Lomadze for regular interconnection [26, sec. 3]. More generally, we describe how to check in Theorem 3.1 whether $Q(M_1)$ is a direct summand of $Q(M)$ and how to construct M_2 if this is the case. The latter algorithm is not complete if the finitely generated module ${}_AM$ is not torsionfree since then $Q(M)$ is not A-finitely generated in general. For the torsionfree modules $U_1 \subseteq U$ in Theorem 3.3 this problem does not arise. We leave it to the future and the younger generation to really implement the algorithms which are only described in words here.

Ideals and, more generally, submodules U of $A^{1\times l}$ are given by finitely many generators. Arbitrary finitely generated A-modules are described by generators and relations in the form $M = A^{1\times l}/U$. Via Gröbner bases also Hom-modules $\mathrm{Hom}_A(M_1, M_2)$ and the annihilators $\mathrm{ann}_A(M)$, $\mathrm{ann}_M(\mathfrak{a})$, $\mathfrak{a} \subseteq A$, and especially $\mathrm{ann}_M(\mathfrak{m}^k)$, $\mathfrak{m} \in \mathrm{Max}(A)$, can be computed. The finite associator $\mathrm{Ass}(M)$ and then also

$$M_{\mathrm{lf}} = \mathrm{tor}_{\mathfrak{T}}(M) = \oplus_{\mathfrak{m} \in \mathrm{Ass}(M) \bigcap \mathrm{Max}(A)} M(\mathfrak{m}),\ M(\mathfrak{m}) := \bigcup_{k=0}^{\infty} \mathrm{ann}_M(\mathfrak{m}^k), \tag{28}$$

and M/M_{lf} can be computed since the ascending sequences of the $\mathrm{ann}_M(\mathfrak{m}^k)$ become stationary. Since $Q(M) = Q(M/M_{\mathrm{lf}})$ one may assume for constructive purposes that $M_{\mathrm{lf}} = \mathrm{tor}_{\mathfrak{T}}(M) = 0$ and $M \subseteq Q(M) \subseteq M_S$ as in Theorem 2.5. In what follows we therefore assume

$$\begin{gathered} M_1 \subseteq M \subseteq Q(M) \subseteq M_S,\ S := \bigcap_{\mathfrak{p} \in \mathrm{Ass}(M)} (A \setminus \mathfrak{p}), \\ Q(M_1) \subseteq Q(M),\ Q(M)/M = \mathrm{tor}_{\mathfrak{T}}(M_S/M). \end{gathered} \tag{29}$$

Algorithm 4.1. *The following algorithm checks for arbitrary dimension $r \geq 2$ whether a subbehavior $\mathcal{B} \subseteq \mathcal{B}_1$ admits a regular almost interconnection and, if so, constructs it. In the situation of Theorem* 3.3 *we assume $U_1 \subseteq U \subseteq A^{1\times l}$ and that U and U_1 are nonzero. We are going to compute $Q(U) = \bigcap_{\mathfrak{p}\in\mathcal{P}_2} U_{\mathfrak{p}} \subseteq A^{1\times l}$; cf.* (21)*. It is clear that infinite intersections cannot, in general, be calculated. The zero ideal is the only prime ideal associated with U since it is torsionfree. Hence $S = A \setminus \{0\}$ in* (29) *and*

$$U \subseteq Q(U) \subset U_S = KU \subseteq K^{1\times l},\ K = F(s),\ Q(U)/U = \mathrm{tor}_{\mathfrak{T}}(KU/U). \tag{30}$$

As submodule of $A^{1\times l} = Q(A^{1\times l})$ the module $Q(U)$ is finitely generated.

We choose a free A-module V with

$$U \subseteq V \subset KU = KV \subseteq KA^{1\times l} = K^{1\times l} \tag{31}$$

and obtain the exact sequence

$$\begin{gathered} 0 \to V/U \xrightarrow{\mathrm{inj}} KU/U \xrightarrow{\mathrm{can}} KU/V \to 0, \text{ and hence} \\ \mathrm{Ass}(KU/U) \subseteq \mathrm{Ass}(V/U) \bigcup \mathrm{Ass}(KU/V) \text{ [4, Chap. IV, sec. 1.2, Prop. 3].} \end{gathered} \tag{32}$$

If $m := \mathrm{rank}(U) := \dim_K(KU)$, then

$$\begin{gathered} V \cong A^{1\times m},\ KV \cong K^{1\times m} \textit{ and } KU/V = KV/V \cong (K/A)^{1\times m}, \textit{ and hence} \\ \mathrm{Ass}(KU/V) = \mathrm{Ass}(K/A) = \{Ap; p \textit{ irreducible}\} \subseteq \mathcal{P}_2. \end{gathered} \tag{33}$$

From the preceding two equations we infer

$$\mathcal{M} := \mathrm{Ass}(KU/U) \bigcap \mathrm{Max}(A) = \mathrm{Ass}(V/U) \bigcap \mathrm{Max}(A). \tag{34}$$

Since V/U is finitely generated, the set $\mathcal{M}$ in (34) *is finite and can be computed. From* (9) *and* (30) *we infer*

$$\begin{gathered} Q(U)/U = \mathrm{tor}_{\mathfrak{T}}(KU/U) = \oplus_{\mathfrak{m}\in\mathcal{M}}(KU/U)(\mathfrak{m}) \textit{ with} \\ (KU/U)(\mathfrak{m}) := \bigcup_{k=0}^{\infty} \mathrm{ann}_{KU/U}(\mathfrak{m}^k). \end{gathered} \tag{35}$$

Notice that for a nonzero ideal $\mathfrak{a}$ and a nonzero element $a \in \mathfrak{a}$ the annihilator

$$\begin{aligned} \mathrm{ann}_{KU/U}(\mathfrak{a}) &:= \left\{\overline{\xi} \in KU/U;\ \mathfrak{a}\xi \subseteq U\right\} = \left\{\overline{\xi} \in a^{-1}U/U;\ \mathfrak{a}\xi \subseteq U\right\} \\ &\cong \{\overline{\eta} \in U/aU;\ \mathfrak{a}\eta \subseteq aU\},\ \overline{\xi} \mapsto \overline{a\xi}, \end{aligned} \tag{36}$$

and especially $\mathrm{ann}_{KU/U}(\mathfrak{m}^k)$ can be computed. Since $Q(U)/U$ is finitely generated there are exponents $e(\mathfrak{m}) \in \mathbb{N}$ with

$$0 = \mathrm{ann}_{KU/U}(\mathfrak{m}^0) \subsetneq \cdots \subsetneq \mathrm{ann}_{KU/U}(\mathfrak{m}^{e(\mathfrak{m})}) = (KU/U)(\mathfrak{m}). \tag{37}$$

These exponents and $(KU/U)(\mathfrak{m})$ can be calculated too. This implies that also generators of

$$Q(U)/U = \mathrm{tor}_{\mathfrak{T}}(KU/U) \textit{ and finally of } Q(U) \tag{38}$$

can be constructively determined. In the same fashion generators of $Q(U_1)$ *can be computed. Assume now that matrices*

$$R \in A^{k\times l} \text{ and } R_1 \in A^{k_1\times l} \text{ with } Q(U_1) = A^{1\times k_1}R_1 \subseteq Q(U) = A^{1\times k}R \tag{39}$$

have been determined. Compute a matrix $R_2 \in A^{g\times k}$ *such that*

$$A^{1\times g}R_2 = \ker\left(A^{1\times k} \overset{\circ R}{\to} Q(U)/Q(U_1) = A^{1\times k}R/A^{1\times k_1}R_1,\ \xi \mapsto \overline{\xi R}\right). \tag{40}$$

According to [26, p. 1077] *the submodule* $Q(U_1)$ *is a direct summand of* $Q(U)$ *if and only if the inhomogeneous linear system*

$$\begin{gathered} R_2R = R_2XR_1 \text{ has a solution } X \in A^{k\times k_1} \text{ and then} \\ Q(U)/Q(U_1) \to Q(U),\ \overline{\xi R} \mapsto \xi(R - XR_1), \end{gathered} \tag{41}$$

is the associated section of the canonical map. If X *exists, we infer*

$$Q(U) = Q(U_1) \oplus A^{1\times k}(R - XR_1), \tag{42}$$

and finally, by Theorems 3.1 *and* 3.3*, the* $\mathfrak{T}$*-isomorphism*

$$+ : U_1 \times U_2 \to U,\ (u_1, u_2) \mapsto u_1 + u_2,\ \text{ with } U_2 := U \bigcap A^{1\times k}(R - XR_1) \tag{43}$$

and the regular almost interconnection

$$\begin{gathered} \mathcal{B} := U^\perp \subseteq \mathcal{B}_1 \bigcap \mathcal{B}_2,\ \mathcal{B}_i := U_i^\perp,\ \text{ with} \\ \mathcal{B}_2 = \left(U \bigcap A^{1\times k}(R - XR_1)\right)^\perp = \mathcal{B} + \{w \in \mathcal{F}^l;\ (R - XR_1) \circ w = 0\}. \end{gathered} \tag{44}$$

Corollary 4.2. *The preceding Algorithm* 4.1 *can be applied to the computation of* $Q(M)$ *and* $Q(M_1)$ *in the situation of Theorem* 3.1 *and Corollary* 3.2 *if* M *and hence also* M_1 *are torsionfree. It thus furnishes a constructive almost direct sum decomposition of a controllable behavior* $\mathcal{B}$ *if this exists.*

We proceed with the situation of Theorem 3.1 and use (29) too. The next theorem gives another equivalent condition to the two equivalent conditions from Theorem 3.1, but the new condition is not fully constructive.

Theorem 4.3. *For the data from* (29) *the following properties are equivalent:*

1. $Q(M_1)$ *is a direct summand of* $Q(M)$ *or there is a submodule* M_2 *of* M *such that* $+ : M_1 \times M_2 \to M$ *is a* $\mathfrak{T}$*-isomorphism.*
2. *There is an* $s \in S$ *and a linear map* $f : M \to M_1$ *such that* $f(x_1) = sx_1$ *for all* $x_1 \in M_1$*, and hence* $f(M_1) = sM_1$*, and* $f(M)/sM_1 \in \mathfrak{C}$*.*

Then also $M_{1,S}$ *is a direct summand of* M_S*.*

For fixed s *the existence of* f *in condition* 2 *can be checked via Gröbner bases again. There is no algorithm, however, which does this for all infinitely many* s*. The method of Zerz and Lomadze* [26] *also enables us to check whether* $M_{1,S}$ *is a direct summand of* M_S *and, if so, to construct a direct complement. However, in general this direct decomposition does not imply a direct decomposition* $Q(M) = Q(M_1) \oplus V$ *as in Algorithm* 4.1*.*

Proof. 2 $\Longrightarrow$ 1: The map $s\circ : M_S \mapsto M_S$ is an isomorphism. Thus f induces the map

$$g := s^{-1}f : M \to s^{-1}M_1 \text{ with } g(x_1) = x_1 \text{ for } x_1 \in M_1 \text{ and}$$
$$g(M)/M_1 \subseteq \mathrm{tor}_{\mathfrak{T}}(s^{-1}M_1/M_1) \subseteq \mathrm{tor}_{\mathfrak{T}}(M_{1,S}/M_1) = Q(M_1)/M_1, \text{ and hence}$$
$$g : M \to Q(M_1) \text{ and } g|M_1 = \mathrm{inj} : M_1 \to Q(M_1).$$

Since $Q(M_1)$ is $\mathfrak{T}$-closed and since $Q(M)/M \in \mathfrak{C}$ the map g admits a unique extension $h : Q(M) \to Q(M_1)$ which also satisfies $h|M_1 = \mathrm{inj}$ and therefore $h|Q(M_1) = \mathrm{id}_{Q(M_1)}$. The preceding arguments hold by means of Lemma 2.3. We conclude $Q(M) = Q(M_1) \oplus \ker(h)$ as asserted. The map $g : M \to s^{-1}M_1 \subseteq M_{1,S}$ with $g|M_1 = \mathrm{inj}$ also induces the map

$$g_S : M_S \to M_{1,S} \text{ with } g_S|M_{1,S} = \mathrm{id}_{M_{1,S}}, \text{ and hence } M_S = M_{1,S} \oplus \ker(g_S).$$

1 $\Longrightarrow$ 2: Let $g : Q(M) \to Q(M_1)$ be a retraction of the injection, i.e., with $g(x_1) = x_1$ for all $x_1 \in Q(M_1)$. This induces

$$g(M) \subseteq Q(M_1) \subseteq M_{1,S} = \bigcup_{s\in S} s^{-1}M_1 \text{ and } g(M)/M_1 \subseteq \mathrm{tor}_{\mathfrak{T}}(M_{1,S}/M_1).$$

Since M is finitely generated and since the submodules $s^{-1}M_1$ of $M_{1,S}$ are directed upwards, there is an $s \in S$ such that

$$g(M) \subseteq s^{-1}M_1,\ g(x_1) = x_1 \text{ for } x_1 \in M_1 \text{ and } g(M)/M_1 \subseteq \mathrm{tor}_{\mathfrak{T}}(s^{-1}M_1/M_1).$$

The map $f := sg : M \to M_1$ then has the asserted properties. □

Corollary 4.4. *With the notation from* (29) *the quotient module $Q(M)$ of a finitely generated module M with $M_{\mathrm{lf}} = 0$ admits the representation*

$$Q(M) = \bigcup_{s\in S} U(s),\ M \subseteq U(s) \subseteq s^{-1}M,\ U(s)/M = \mathrm{tor}_{\mathfrak{T}}(s^{-1}M/M).$$

Each single $U(s)$ can be computed, but the infinite directed union is not finitely generated in general and cannot be calculated.

5. Stable finite-dimensional systems. In this section we are going to talk about stable systems and therefore assume that F is the field $\mathbb{C}$ of complex numbers. To motivate the following considerations consider the $\mathbb{C}$-one-dimensional, unstable differential behavior

$$\mathcal{B} := \{y \in \mathrm{C}^{\infty}(\mathbb{R}, \mathbb{C}); (s-1)\circ y = 0\} = \mathbb{C}e^t.$$

This example shows that the consideration of a behavior up to a $\mathbb{C}$-finite-dimensional one as in the preceding sections may be misleading if this finite-dimensional part is unstable. In stability and stabilization theory it is therefore customary to neglect only those autonomous systems which are stable in a suitable sense, for instance, asymptotically stable in the standard one-dimensional theory. Feedback stabilization of multidimensional input/output systems was treated with this philosophy in [14]. In this short section we describe how the theory of the preceding sections can be extended to the case where only stable finite-dimensional behaviors are neglected. It turns out that the localization theory is very versatile and can be easily adapted to this situation.

The assumptions of section 2 remain in force. As in [14, sec. 5, eq. (70)] we additionally choose a disjoint decomposition

$$\mathbb{C}^r := \Lambda_1 \uplus \Lambda_2 \tag{45}$$

into a *stable part* Λ_1 and an *unstable part* Λ_2. The standard choice in the continuous case is $\Lambda_2 := \{z \in \mathbb{C};\ \Re(z) \geq 0\}^r$. For $r = 1$ the set Λ_1 consists of the complex numbers with negative real part which is customarily used for continuous stabilization theory. Multidimensional stability and stabilization with respect to (45) have been discussed in [14]. In the complex continuous case the decomposition (45) implies the direct sum decomposition

$$\mathcal{F}_1 = \mathcal{F}_{\mathrm{lf}} = \mathcal{F}_{\mathrm{lf},1} \oplus \mathcal{F}_{\mathrm{lf},2},\ \mathcal{F}_{\mathrm{lf},i} = \oplus_{\lambda \in \Lambda_i} \mathbb{C}[t] e^{\lambda \bullet t},\ i = 1, 2, \tag{46}$$

of the space of polynomial-exponential functions. For the real and the discrete cases analogous decompositions hold. These decompositions, in turn, furnish the direct decomposition

$$\mathcal{F} = \mathcal{F}_{\mathrm{lf}} \oplus \mathcal{F}_2 = \mathcal{F}_1^\Lambda \oplus \mathcal{F}_2^\Lambda,\ \mathcal{F}_1^\Lambda := \mathcal{F}_{\mathrm{lf},1},\ \mathcal{F}_2^\Lambda = \mathcal{F}_{\mathrm{lf},2} \oplus \mathcal{F}_2. \tag{47}$$

The map

$$\mathbb{C}^r \cong \mathrm{Max}(A),\ \lambda \mapsto \mathfrak{m}_\lambda := \sum_{k=1}^r A(s_k - \lambda_k), \tag{48}$$

is bijective. The decomposition $\mathcal{F} := \mathcal{F}_1^\Lambda \oplus \mathcal{F}_2^\Lambda$ then induces the disjoint decomposition

$$\begin{gathered} \mathrm{Spec}(A) = \mathcal{P}_1^\Lambda \uplus \mathcal{P}_2^\Lambda \text{ with } \mathcal{P}_1^\Lambda := \mathrm{Ass}(\mathcal{F}_1^\Lambda) = \{\mathfrak{m}_\lambda;\ \lambda \in \Lambda_1\} \subseteq \mathrm{Max}(A) \\ \text{and } \mathcal{P}_2^\Lambda := \mathrm{Ass}(\mathcal{F}_2^\Lambda) = (\mathrm{Spec}(A) \setminus \mathrm{Max}(A)) \uplus \{\mathfrak{m}_\lambda;\ \lambda \in \Lambda_2\}. \end{gathered} \tag{49}$$

A locally finite (torsion) module M is called *stable* with respect to the decomposition (45) if $\mathrm{Ass}(M) \subseteq \mathcal{P}_1^\Lambda$. If $M = A^{1\times p}/U$ is stable and $\mathbb{C}$-finite-dimensional, its associated autonomous behavior $\mathcal{B} \cong \mathrm{Hom}_A(M, \mathcal{F})$ consists of polynomial-exponential trajectories with exponents in Λ_1 [13, eq. (38)]; indeed

$$\mathcal{B} = \oplus_{\lambda \in \Lambda_1,\ \mathfrak{m}_\lambda \in \mathrm{Ass}(M)} \mathcal{B}(\mathfrak{m}_\lambda),\ \mathcal{B}(\mathfrak{m}_\lambda) \subseteq \mathbb{C}[t]^p e^{\lambda \bullet t}. \tag{50}$$

For $r = 1$ and $\Lambda_1 = \{z \in \mathbb{C};\ \Re(z) < 0\}$ this signifies that $\mathcal{B}$ is asymptotically stable.

The Serre subcategory $\mathfrak{C}$ of locally finite modules and its associated Gabriel topology is now replaced by the subcategory and Gabriel topology

$$\begin{gathered} \mathfrak{C}^\Lambda := \{C \in \mathrm{Mod}_A;\ \mathrm{Hom}_A(C, \mathcal{F}_2^\Lambda) = 0\} = \{C \in \mathrm{Mod}_A;\ \forall \mathfrak{p} \in \mathcal{P}_2^\Lambda : C_\mathfrak{p} = 0\} \subseteq \mathfrak{C}, \\ \mathfrak{T}^\Lambda := \{\mathfrak{a} \subseteq A;\ A/\mathfrak{a} \in \mathfrak{C}^\Lambda\} \subseteq \mathfrak{T}, \\ \mathfrak{C}^\Lambda = \{C \in \mathrm{Mod}_A;\ \forall x \in C : \mathrm{ann}_A(x) \in \mathfrak{T}^\Lambda\} = \{C \in \mathrm{Mod}_A;\ \mathrm{Ass}(C) \subseteq \mathcal{P}_1^\Lambda\}. \end{gathered} \tag{51}$$

$\mathfrak{T}^\Lambda$-torsion, $\mathfrak{T}^\Lambda$-torsionfree, and $\mathfrak{T}^\Lambda$-closed modules, the category $\mathrm{Mod}_{A,\mathfrak{T}^\Lambda}$, and the quotient functor $Q^\Lambda : \mathrm{Mod}_A \to \mathrm{Mod}_{A,\mathfrak{T}^\Lambda}$ are defined in analogy to the case of the topology $\mathfrak{T}$ and have the corresponding properties. The $\mathfrak{T}^\Lambda$-torsion radical $\mathrm{tor}_{\mathfrak{T}^\Lambda}(M)$ is the largest $\mathfrak{T}^\Lambda$-torsion submodule of M and contained in $M_{\mathrm{lf}} = \mathrm{tor}_{\mathfrak{T}}(M)$. Every $\mathfrak{T}$-closed module is also $\mathfrak{T}^\Lambda$-closed since $\mathfrak{T}^\Lambda \subseteq \mathfrak{T}$. Therefore every A-module M gives rise to the commutative diagram with exact rows,

(52)
$$\begin{tikzcd} 0 \arrow[r] & \mathrm{tor}_{\mathfrak{T}^\Lambda}(M) \arrow[r, "\mathrm{inj}"] \arrow[d, phantom, "\bigcap"] & M \arrow[r, "\eta^\Lambda"] \arrow[d, phantom, "\|"] & Q^\Lambda(M) = Q^\Lambda(M/\mathrm{tor}_{\mathfrak{T}^\Lambda}(M)) \arrow[d, "g = Q^\Lambda(\eta)"] \\ 0 \arrow[r] & \mathrm{tor}_{\mathfrak{T}}(M) \arrow[r, "\mathrm{inj}"] & M \arrow[r, "\eta"] & Q(M) = Q^\Lambda Q(M), \end{tikzcd}$$

where $g = Q^\Lambda(\eta)$ is the unique A-linear map with $g\eta^\Lambda = \eta$.

Mutatis mutandis equations (14)–(22), Lemma 2.3, Theorems 2.4, 2.5, 3.1, and 3.3, and Algorithm 4.1 hold when $\mathcal{F}_1, \mathcal{F}_2, \mathcal{P}_1, \mathcal{P}_2$ are replaced by $\mathcal{F}_1^\Lambda, \mathcal{F}_2^\Lambda, \mathcal{P}_1^\Lambda, \mathcal{P}_2^\Lambda$. We thus obtain a theory of behaviors up to $\mathbb{C}$-finite-dimensional, Λ-stable ones.

Remark 5.1. The localization theory can be applied to more general decompositions $\mathrm{Spec}(A) = \mathcal{Q}_1 \uplus \mathcal{Q}_2$ with the property that $\mathfrak{p}_1 \subset \mathfrak{p}_2 \in \mathcal{Q}_2$ implies $\mathfrak{p}_1 \in \mathcal{Q}_2$, for instance, for given n with $0 < n \le r$,

$$\mathcal{Q}_1 := \{\mathfrak{p} \in \mathrm{Spec}(A);\ \dim(A/\mathfrak{p}) < n\},\ \mathcal{Q}_2 := \{\mathfrak{p} \in \mathrm{Spec}(A);\ \dim(A_\mathfrak{p}) \le r - n\}.$$

Here we used $\dim(A_\mathfrak{p}) + \dim(A/\mathfrak{p}) = r$. In particular,

$$\mathcal{P}_1 = \mathrm{Max}(A) = \{\mathfrak{p} \in \mathrm{Spec}(A);\ \dim(A/\mathfrak{p}) < 1\},$$
$$\mathcal{P}_2 = \mathrm{Spec}(A) \setminus \mathrm{Max}(A) = \{\mathfrak{p};\ \dim(A_\mathfrak{p}) \le r - 1\}.$$

The choice

$$\mathcal{Q}_1 := \{\mathfrak{p} \in \mathrm{Spec}(A);\ \dim(A/\mathfrak{p}) < r - 1\},$$
$$\mathcal{Q}_2 := \{\mathfrak{p} \in \mathrm{Spec}(A);\ \dim(A_\mathfrak{p}) \le 1\}$$

leads to the pseudozero modules and pseudoisomorphisms from [4, Chap. VII, sec. 4.4, Defs. 2 and 3].

Remark 5.2. In [14] the localization theory was applied to the Serre subcategory and Gabriel topology

$$\begin{aligned} &\mathfrak{C}^{\mathrm{IO},\Lambda} := \{C \in \mathrm{Mod}_A;\ \forall \lambda \in \Lambda_2 : C_{\mathfrak{m}_\lambda} = 0\} \supset \\ &\mathfrak{C}^\Lambda = \left\{C \in \mathfrak{C}^{\mathrm{IO},\Lambda};\ C \text{ locally finite}\big)\right\} \text{ and } \mathfrak{T}^{\mathrm{IO},\Lambda} := \left\{\mathfrak{a} \subseteq A;\ A/\mathfrak{a} \in \mathfrak{C}^{\mathrm{IO},\Lambda}\right\}. \end{aligned} \tag{53}$$

The corresponding localization functor $Q^{\mathrm{IO},\Lambda}$ satisfies [14, Lem. 3.4]

$$Q^{\mathrm{IO},\Lambda}(A) = A_T \text{ with } T := \{t \in A;\ \forall \lambda \in \Lambda_2 : t(\lambda) \neq 0\}. \tag{54}$$

A finitely generated (torsion) module $M \in \mathfrak{C}^{\mathrm{IO},\Lambda}$ gives rise to an autonomous behavior $\mathcal{B}^0 := \mathrm{Hom}_A(M, \mathcal{F})$ which we call Λ-stable or T-stable. In [14] we mainly discussed its part $\mathrm{Hom}_A(M, \mathcal{F}_{\mathrm{lf}})$ of polynomial-exponential trajectories. Theorems 3.1 and 3.3 also hold for this Serre subcategory and associated quotient module and give rise to a theory of regular $\mathfrak{T}^{\mathrm{IO},\Lambda}$-interconnections up to T-stable autonomous behaviors. Since $\mathfrak{C}^{\mathrm{IO},\Lambda}$ is much bigger than $\mathfrak{C}^\Lambda$ from (51) and contains many nonlocally finite modules, there are many more associated regular $\mathfrak{T}^{\mathrm{IO},\Lambda}$- than $\mathfrak{T}^\Lambda$-interconnections.

The torsion radical $\mathrm{tor}_{\mathfrak{T}^{\mathrm{IO},\Lambda}}(M)$ is given by the infinite intersection [14, Lemma 3.5]

$$\begin{aligned}\mathrm{tor}_{\mathfrak{T}^{\mathrm{IO},\Lambda}}(M) &= \bigcap_{\lambda\in\Lambda_2} \ker\left(M \xrightarrow{\mathrm{can}} M_{\mathfrak{m}_\lambda},\ x \mapsto \frac{x}{1}\right)\\ &= \{x \in M;\ \forall\lambda \in \Lambda_2 \exists t_\lambda \in A \text{ with } t_\lambda(\lambda) \neq 0 \text{ and } t_\lambda x = 0\}\end{aligned} \tag{55}$$

and cannot be easily computed in general. The quotient module $Q^{\mathrm{IO},\Lambda}(U)$ of a finitely generated torsionfree module $U \subseteq A^{1\times l}$ is also given by the infinite intersection [14, Lem. 3.6]

$$Q^{\mathrm{IO},\Lambda}(U) = A_T^{1\times l} \bigcap \bigcap_{\lambda\in\Lambda_2} U_{\mathfrak{m}_\lambda} \tag{56}$$

and is also hard to compute in general. Define, however, $M := A^{1\times l}/U$ and assume that all modules $M_{\mathfrak{m}_\lambda} = A_{\mathfrak{m}_\lambda}^{1\times l}/U_{\mathfrak{m}_\lambda}$, $\lambda \in \Lambda_2$, are torsionfree. Then $Q^{\mathrm{IO},\Lambda}(U)$ can be computed as [14, Thm. 5.14]

$$Q^{\mathrm{IO},\Lambda}(U) = U_{\mathrm{cont},T} \subseteq A_T^{1\times l}, \text{ where } U_{\mathrm{cont}}/U = \mathrm{tor}(A^{1\times l}/U) \tag{57}$$

is the torsion submodule of M. If even M is torsionfree, i.e., if the behavior $\mathrm{Hom}_A(M, \mathcal{F})$ is controllable, then $Q^{\mathrm{IO},\Lambda}(U) = U_T$ is the usual quotient module and its A_T-generators are the A-generators of U.

Hence, if in Theorem 3.3 the given behaviors $\mathcal{B} \subseteq \mathcal{B}_1$ are controllable, matrices

$$\begin{gathered} R \in A^{k\times l},\ R_1 \in A^{k_1\times l} \text{ with } U = A^{1\times k}R,\ U_1 = A^{1\times k_1}R_1 \text{ and}\\ Q(U) = U_T = A_T^{1\times k}R,\ Q(U_1) = U_{1,T} = A_T^{1\times k_1}R_1\end{gathered} \tag{58}$$

are given. Hence $\mathcal{B} \subseteq \mathcal{B}_1$ admit a regular $\mathfrak{T}^{\mathrm{IO},\Lambda}$-interconnection if and only if equations (41) have a solution $X \in A_T^{k\times k_1}$. But there is yet no algorithm which solves *inhomogeneous* systems of linear equations in the quotient ring

$$A_T,\ T = \{t \in A;\ \forall\lambda \in \Lambda_2 : t(\lambda) \neq 0\},$$

as pointed out in [14, Rem. 5.10]; cf. also [7] and [25].

Acknowledgments. I appreciate the referees' work in my interest. I thank D. Napp Avelli and S. Shankar for pointing out that the behavior interconnection in Theorem 3.3 is indeed regular. Although this is a simple observation, I had missed it.

REFERENCES

[1] M. Bisiacco and M. E. Valcher, *A note on the direct decomposition of two-dimensional behaviors*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 48 (2001), pp. 490–494.

[2] M. Bisiacco and M. E. Valcher, *Two-dimensional behavior decompositions with finite-dimensional intersection: A complete characterization*, Multidimens. Syst. Signal Process., 16 (2005), pp. 335–354.

[3] V. D. Blondel and A. Megretski, eds., *Unsolved Problems in Mathematical Systems and Control Theory*, Princeton University Press, Princeton, NJ, 2004.

[4] N. Bourbaki, *Commutative Algebra*, Addison–Wesley, Reading, MA, 1972.

[5] P. Gabriel, *Des catégories abéliennes*, Bull. Soc. Math. France, 90 (1962), pp. 323–448.

[6] K. Galkowski and J. Wood, eds., *Multidimensional Signals, Circuits and Systems*, Taylor and Francis, London, 2001.

[7] Z. Lin, *Output feedback stabilizability and stabilization of linear n-D systems*, in Multidimensional Signals, Circuits and Systems, K. Galkowski and J. Wood, eds., Taylor and Francis, London, 2001, pp. 59–76.

[8] H. Matsumura, *Commutative Ring Theory*, Cambridge University Press, Cambridge, UK, 1986.

[9] D. Napp Avelli, *An Algebraic Approach to Multidimensional Behaviors*, Ph.D. thesis, Groningen, 2007.

[10] D. Napp Avelli, *Almost direct sum decomposition and implementation of* 2*D behaviors*, Math. Control Signals Systems, to appear.

[11] U. Oberst, *Multidimensional constant linear systems*, Acta Appl. Math., 20 (1990), pp. 1–175.

[12] U. Oberst, *Variations on the fundamental principle for linear systems of partial differential and difference equations with constant coefficients*, Appl. Algebra Engrg. Comm. Comput., 6 (1995), pp. 211–243.

[13] U. Oberst, *Finite dimensional systems of partial differential or difference equations*, Adv. Appl. Math., 17 (1996), pp. 337–356.

[14] U. Oberst, *Stability and stabilization of multidimensional input/output systems*, SIAM J. Control Optim., 45 (2006), pp. 1467–1507.

[15] U. Oberst and F. Pauer, *The constructive solution of linear systems of partial difference and differential equations with constant coefficients*, Multidimens. Systems Signal Process., 12 (2001), pp. 253–308.

[16] J. W. Polderman and J. C. Willems, *Introduction to Mathematical Systems Theory*, Springer-Verlag, New York, 1998.

[17] P. Rocha, *Feedback control of multidimensional behaviors*, Systems Control Lett., 45 (2002), pp. 207–215.

[18] P. Rocha and J. Wood, *Trajectory control and interconnection of* 1*D and nD systems*, SIAM J. Control Optim., 40 (2001), pp. 107–134.

[19] S. Shankar, *Can one control the vibrations of a drum?*, Multidimens. Systems Signal Process., 11 (2000), pp. 67–81.

[20] S. Shankar, *The lattice structure of behaviors*, SIAM J. Control Optim., 39 (2001), pp. 1817–1832.

[21] B. Stenström, *Rings of Quotients*, Springer-Verlag, Berlin, 1975.

[22] M. E. Valcher, *On the decomposition of two-dimensional behaviors*, Multidimens. Systems Signal Process., 11 (2000), pp. 49–65.

[23] J. C. Willems, *Paradigms and puzzles in the theory of dynamical systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 259–294.

[24] J. C. Willems, *On interconnections, control and feedback*, IEEE Trans. Automat. Control, 42 (1997), pp. 458–472.

[25] L. Xu, Z. Lin, J.-Q. Ying, O. Saito, and Y. Anazawa, *Open problems in control of linear discrete multidimensional systems*, in Unsolved Problems in Mathematical Systems and Control Theory, V. D. Blondel and A. Megretski, eds., Princeton University Press, Princeton, NJ, 2004, pp. 221–228.

[26] E. Zerz and V. Lomadze, *A constructive solution to interconnection and decomposition problems with multidimensional behaviors*, SIAM J. Control Optim., 40 (2001), pp. 1072–1086.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 2, pp. 481–520

© 2009 Society for Industrial and Applied Mathematics

UTILITY MAXIMIZATION WITH HABIT FORMATION: DYNAMIC PROGRAMMING AND STOCHASTIC PDEs[*]

NIKOLAOS ENGLEZOS[†] AND IOANNIS KARATZAS[†]

Abstract. This paper studies the habit-forming preference problem of maximizing total expected utility from consumption net of the *standard of living*, a weighted average of past consumption. We describe the effective state space of the corresponding optimal wealth and standard of living processes, identify the associated value function as a generalized utility function, and exploit the interplay between dynamic programming and Feynman–Kac results via the theory of random fields and stochastic partial differential equations (SPDEs). The resulting value random field of the optimization problem satisfies a nonlinear, backward SPDE of parabolic type, widely referred to as the *stochastic Hamilton–Jacobi–Bellman equation.* The dual value random field is characterized further in terms of a backward parabolic SPDE which is *linear*. Progressively measurable versions of stochastic feedback formulae for the optimal portfolio and consumption choices are obtained as well.

Key words. habit formation, generalized utility function, random fields, stochastic backward partial differential equations, feedback formulae, stochastic Hamilton–Jacobi–Bellman equation

AMS subject classifications. Primary, 93E20, 60H15, 91B28; Secondary, 91B16, 35R60

DOI. 10.1137/070686998

1. Introduction.

An important question in financial mathematics is to explain the effect of past consumption patterns on current and future economic decisions. A useful tool in this effort has been the concept of *habit formation*: an individual who consumes portions of his wealth over time is expected to develop habits which will have a decisive impact on his subsequent consumption behavior. Employed in a wide variety of economic applications, habit formation was in turn considered by several authors in the classical utility optimization problem (e.g., Sundaresan (1989), Constantinides (1990), Detemple and Zapatero (1991, 1992), Heaton (1993), Chapman (1998), Schroder and Skiadas (2002)).

The present paper studies portfolio/consumption optimization under habit formation in the complete market model of Detemple and Zapatero (1992). We adopt nonseparable von Neumann–Morgenstern preferences over a given time horizon $[0,T]$ and study the stochastic control problem of maximizing total expected utility

$$E\int_0^T u\big(t, c(t) - z(t;c)\big)\,dt \tag{1.1}$$

from consumption $c(\cdot)$ in excess of a *standard of living* (or "habit index") $z(\cdot;c)$. This is a weighted linear average of past consumption, given by

$$z(t;c) \triangleq z\,e^{-\int_0^t \alpha(v)dv} + \int_0^t \delta(s)\,e^{-\int_s^t \alpha(v)dv} c(s)\,ds, \qquad 0 \le t \le T, \tag{1.2}$$

with $z \ge 0$ and nonnegative stochastic coefficients $\alpha(\cdot)$, $\delta(\cdot)$. Moreover, by assuming infinite marginal utility at zero, i.e., $u'(t,0^+) = \infty$, we force consumption never to

[*] Received by the editors March 30, 2007; accepted for publication (in revised form) September 2, 2008; published electronically February 13, 2009. This paper draws on results obtained in the first author's doctoral dissertation at Columbia University, February 2007. This work was supported in part by the National Science Foundation under grant NSF-DMS-06-01774.

http://www.siam.org/journals/sicon/48-2/68699.html

[†] Department of Mathematics, Columbia University, New York, NY 10027 (negglez@math.columbia.edu, ik@math.columbia.edu).

fall below the current level of standard of living, thus triggering the development of "addictive" consumption patterns: the agent is constantly "forced" to consume more than he used to in the past. At $t = 0$ the assumption $u'(t, 0^+) = \infty$ postulates the condition $x > wz$, specifying the *wedge* $\mathcal{D}$ of Assumption 4.1 as the *domain of acceptability* for the initial wealth x and initial standard of living z. The quantity w stands for the cost, per unit of standard of living, of the *subsistence consumption*: the consumption policy that matches the standard of living exactly, at all times.

Existence of an optimal portfolio/consumption pair is proved in Detemple and Zapatero (1992) by establishing a recursive linear stochastic equation for the properly normalized marginal utility. In order to set up the mathematical background needed for further analysis, we present a brief formulation of their solution. Our contribution starts by characterizing the *effective state space* of the corresponding optimal wealth $X_0(\cdot)$ and standard of living $z_0(\cdot)$ processes as the *random wedge* $\mathcal{D}_t$ (cf. (5.18)) determined by the evolution $\mathcal{W}(t)$ of w as a random process. This result reveals the stochastic evolution of the imposed condition $x > wz$ over time, in the sense that $X_0(t) > \mathcal{W}(t)z_0(t)$ for all $t \in [0, T)$, and motivates the study of the *dynamic* aspects of our stochastic control problem. Thus, we define the value function V of the optimization problem as a mapping that depends on both x and z. Considering the latter as a pair of variables running on $\mathcal{D}$, we classify V in a broad family of utility functions; in fact, $V(\cdot, z)$ and the utility function $u(t, \cdot)$ exhibit similar analytic properties. This is carried out through the *convex dual* of the value function, defined in (5.24), in conjunction with differential techniques developed in Rockafellar (1970).

In order to describe quantitatively the dependence of the agent's optimal investment $\pi_0(\cdot)$ on his wealth $X_0(\cdot)$ and standard of living $z_0(\cdot)$, Detemple and Zapatero (1992) restrict the utility function to have either the logarithmic $u(t, x) = \log x$ or the power $u(t, x) = x^p/p$ form for a model with nonrandom coefficients. Driven by ideas of dynamic programming, we pursue such formulae for the optimal policies, where now u can be an arbitrary utility function and the model coefficients may be random in general. Classical dynamic programming techniques are, however, inadequate for the analysis of non-Markovian models. On the other hand, the dynamic evolution of domain $\mathcal{D}$, represented by the stochastically evolving wedges $\mathcal{D}_t$, hints that the *basic principles* of dynamic programming might be applicable in more general settings as well. Indeed, Peng (1992) considered a stochastic control problem with stochastic coefficients and made use of Bellman's optimality principle to formulate an associated *stochastic* Hamilton–Jacobi–Bellman equation. The discussion in that paper was formal, due to insufficient regularity of the value function. The present paper culminates with *an explicit application and validation of Peng's ideas for utility maximization.*

Since stock prices and the money-market price are not necessarily Markov processes, we are now required to work with conditional expectations; these take into account the market history up to the present and thereby lead to the consideration of *random fields.* In this context, an important role is played by certain linear, *backward* parabolic *stochastic* partial differential equations which characterize the resulting random fields as their *unique adapted* solutions; in other words, *adapted* versions of *stochastic* Feynman–Kac formulae are established.

Under reasonable assumptions on the utility preferences, the adapted *value random field* of the stochastic control problem solves, in the classical sense, a nonlinear, backward *stochastic* partial differential equation of parabolic type. Namely, the value random field possesses sufficient smoothness, such that all of the spatial derivatives involved in the equation exist almost surely. This equation is the stochastic Hamilton–Jacobi–Bellman equation one would expect, according to the program of

Peng (1992), and is derived from two *linear* Cauchy problems, which admit *unique* solutions subject to certain regularity conditions. Apart from the classical linear/quadratic case discussed in Peng (1992), and to the best of our knowledge, *this work is the first to illustrate explicitly, directly, and completely the role of backward stochastic partial differential equations (SPDEs) in the study of stochastic control problems in any generality*; see Remarks 7.5 and 7.6 in this respect.

We also characterize the *dual* value random field as the unique adapted solution of a *linear,* parabolic backward SPDE. We conclude by deriving *stochastic* "feedback formulae" for the optimal portfolio-consumption decisions, in terms of the pair consisting of the current level of wealth and standard of living. In the special case of deterministic coefficients, these formulae establish this pair as a *sufficient statistic* for the optimal investment and consumption actions of an economic agent in this market.

Preview. Sections 2–5 introduce the market model and study the portfolio-consumption stochastic control problem under habit formation. Section 6 investigates the interrelation of dynamic programming with the theory of SPDEs, which establishes the optimal policies in "feedback form". Section 7 develops the stochastic Hamilton–Jacobi–Bellman equation satisfied by the value random field and provides an equivalent characterization for its dual. A stochastic Feynman–Kac formula is presented in section 8 (appendix), and conclusions follow in section 9.

Literature overview. Duality methods in stochastic control were introduced by Bismut (1973) and elaborated further in Xu (1990) and Karatzas and Shreve (1998). Detemple and Zapatero (1991, 1992) employ martingale methods (Cox and Huang (1989), Karatzas (1989), Karatzas, Lehockzy, and Shreve (1987), and Pliska (1986)) to derive a closed-form solution for the optimal consumption policy, denoted by $c_0(\cdot)$. They also provide insights about the structure of the optimal portfolio investment $\pi_0(\cdot)$ that finances the policy $c_0(\cdot)$, via an application of the Clark (1970)/Ocone and Karatzas (1991) formula.

Detemple and Karatzas (2003) explored more recently a case of nonaddictive habits. In particular, their optimization problem deals with a utility function $u{:}[0,T]\times\mathbb{R}\to\mathbb{R}$ (e.g., of exponential type) where the marginal utility at zero is finite and the "addiction" condition (4.2) is removed. Nonetheless, the natural constraint of a nonnegative consumption plan remains intact in the model. The existence of an optimal pair was demonstrated and the optimal consumption process was provided in closed form, in terms of an endogenously determined stopping time, after which the nonnegativity constraint on consumption ceases to be binding.

The use of dynamic programming techniques on stochastic control problems originated with Merton (1969, 1971), who obtained closed-form solutions in the special case of constant coefficients for models without habit formation. The infinite-horizon case was generalized by Karatzas et al. (1986). Karatzas, Lehoczky, and Shreve (1987) coupled martingale methods with convexity methods to allow random, adapted model coefficients for general preferences; nonetheless, they reinstated the Markovian framework with constant coefficients to obtain the optimal portfolio in closed form. A study on the case of deterministic coefficients in markets without habits can be found in Karatzas and Shreve (1998).

"Pathwise" stochastic control problems were studied recently by Lions and Souganidis (1998a, 1998b), who proposed a new notion of *stochastic* viscosity solutions for the associated fully nonlinear *stochastic* Hamilton–Jacobi–Bellman equations. In two subsequent papers, Buckdahn and Ma (2001a, 2001b) employ a Doss–Sussmann-type transformation to extend this notion in a "pointwise" manner and obtain accordingly existence and uniqueness results for similar SPDEs. A problem of "pathwise" stochastic optimization, that emerges from mathematical finance and concerns the dependence on the paths of an exogenous noise, is considered by Buckdahn and Ma (2007).

Results concerning the existence, uniqueness, and regularity of adapted solutions to SPDEs of the type considered in the present paper were obtained in Ma and Yong (1997, 1999). Kunita (1990) contains a systematic study of semimartingales with spatial parameters, including the derivation of the generalized Itô–Kunita–Wentzell formula that is put to significant use throughout our analysis.

The setting of (1.1), (1.2) represents only one of many possible ways in which to capture effects of habit formation and/or intertemporal substitution of preferences. We refer the reader to Hindy and Huang (1993), Hindy, Huang, and Kreps (1992), and Bank and Riedel (2000, 2001) for alternative approaches using other functionals of past and present consumption, as well as to the references in these papers and to the survey by Bank and Föllmer (2003).

Appendix on notation. The paper is quite heavy with notation, so here is a partial list for ease of reference:

- $u(\cdot), \widetilde{u}(\cdot), I(\cdot)$: a utility function, its convex dual, and the inverse function of its derivative $u'(\cdot)$, respectively; see section 3.
- $Z(\cdot), Z^t(\cdot)$: exponential (local) martingales; (2.6) and (6.4), respectively.
- $H(\cdot), H^t(\cdot)$: state price density processes; (2.8) and (6.4), respectively.
- $\Gamma(\cdot), \Gamma^t(\cdot)$: state price density processes adjusted for habit formation; (5.1) and (6.5), respectively.
- $V(x,z), \mathrm{V}(t,x,z)$: the value function of (4.8) and the value random field of (7.4), respectively.
- $\widetilde{V}(y), \widetilde{\mathrm{V}}(t,y)$: the convex duals of the value function $V(\cdot,\cdot)$ and of the random field $\mathrm{V}(t,\cdot,\cdot)$ as above; they appear in (5.24) and (7.23), respectively.
- w, $\mathcal{W}(\cdot)$, and $w(\cdot)$: the marginal costs of subsistence consumption in (4.6), (5.12), in (5.19), and in (6.6), respectively.
- $\mathcal{X}(\cdot)$, $G(\cdot)$ and $\mathfrak{X}(t,\cdot)$, $\mathfrak{G}(t,\cdot)$: the auxiliary functions of (5.3), (5.17) and the auxiliary random fields of (6.13), (7.7), respectively.
- $\mathcal{Y}(\cdot)$, $\mathfrak{Y}(t,\cdot)$: the inverses of the function $\mathcal{X}(\cdot)$ and of the random field $\mathfrak{X}(t,\cdot)$.
- $c_0(\cdot)$, $\pi_0(\cdot)$, and $X_0(\cdot)$: the optimal consumption, portfolio, and wealth processes of (5.10), (5.14), and (5.13), respectively.
- $\mathcal{A}'(x,z)$, $\mathcal{A}'(t,x,z)$: classes of admissible portfolio/consumption process pairs; Definition 4.2 and (7.5), respectively.
- $\mathcal{B}'(x,z)$: class of admissible consumption processes; Definition 4.3.
- $C(t,x,z)$, $\Pi(t,x,z)$: the optimal consumption and portfolio random fields of (6.20) and (6.21), respectively.

The following notation will also be in use throughout the paper:

For any integer $k \geq 0$, let

- $C^k(\mathbb{R}^n;\mathbb{R}^d)$ denote the set of functions from $\mathbb{R}^n$ to $\mathbb{R}^d$ that are continuously differentiable up to order k.

Consider a probability space $(\Omega, \mathcal{F}, P)$ endowed with a filtration $\mathbb{F}$. Then for any $1 \leq p \leq \infty$, any Banach space $\mathbb{X}$ with norm $\|\cdot\|_{\mathbb{X}}$, and any sub-σ-algebra $\mathcal{G} \subseteq \mathcal{F}$, let

- $\mathbb{L}^p_{\mathcal{G}}(\Omega;\mathbb{X})$ denote the set of all $\mathbb{X}$-valued, $\mathcal{G}$-measurable random variables $\mathbf{X}$ such that $E\|\mathbf{X}\|^p_{\mathbb{X}} < \infty$;
- $\mathbb{L}^p_{\mathbb{F}}(0,T;\mathbb{X})$ denote the set of all $\mathbb{F}$-progressively measurable, $\mathbb{X}$-valued processes $\mathbf{X} : [0,T] \times \Omega \to \mathbb{X}$ such that $\int_0^T \|\mathbf{X}(t)\|^p_{\mathbb{X}}\, dt < \infty$ almost surely;

- $\mathbb{L}^p_{\mathbb{F}}(0,T;\mathbb{L}^p(\Omega;\mathbb{X}))$ denote the set of all $\mathbb{F}$-progressively measurable, $\mathbb{X}$-valued processes $\mathbf{X}:[0,T]\times\Omega\to\mathbb{X}$ such that $\int_0^T E\|\mathbf{X}(t)\|^p_{\mathbb{X}}\,dt<\infty$;
- $C_{\mathbb{F}}([0,T];\mathbb{X})$ denote the set of all continuous, $\mathbb{F}$-adapted processes $\mathbf{X}(\cdot,\omega):[0,T]\to\mathbb{X}$ for P-a.e. $\omega\in\Omega$.

We shall define similarly the set $C_{\mathbb{F}}([0,T];\mathbb{L}^p(\Omega;\mathbb{X}))$, and let $\mathbb{R}^+$ stand for the positive real numbers.

2. The model. We adopt a model for the financial market $\mathcal{M}_0$ which consists of one riskless asset (money market) with price $S_0(t)$ given by

$$dS_0(t)=r(t)S_0(t)dt,\qquad S_0(0)=1, \tag{2.1}$$

and m risky securities (stocks) with prices per share $\{S_i(t)\}_{1\le i\le m}$, satisfying the stochastic differential equations

$$dS_i(t)=S_i(t)\left[b_i(t)dt+\sum_{j=1}^d\sigma_{ij}(t)\,dW_j(t)\right],\quad i=1,\dots,m. \tag{2.2}$$

Here $W(\cdot)=(W_1(\cdot),\dots,W_d(\cdot))^*$ is a d-dimensional Brownian motion on a probability space $(\Omega,\mathcal{F},P)$ and $\mathbb{F}=\{\mathcal{F}(t);\ 0\le t\le T\}$ will denote the P-augmentation of the Brownian filtration $\mathcal{F}^W(t)\triangleq\sigma(W(s);\ s\in[0,t])$. We assume that $d\ge m$; i.e., the number of sources of uncertainty in the model is at least as large as the number of stocks available for investment. All processes encountered in this paper are defined on a fixed, finite time horizon $[0,T]$, where T is the *terminal time.*

The interest rate $r(\cdot)$, as well as the instantaneous rate of return vector $b(\cdot)=(b_1(\cdot),\dots,b_m(\cdot))^*$ and the volatility matrix $\sigma(\cdot)=\{\sigma_{ij}(\cdot)\}_{1\le i\le m,1\le j\le d}$, are taken to be $\mathbb{F}$-progressively measurable random processes and to satisfy

$$\int_0^T\|b(t)\|dt<\infty,\qquad\int_0^T|r(t)|dt\le\varrho \tag{2.3}$$

almost surely for some given real constant $\varrho>0$. It will be assumed that $\sigma(\cdot)$ is bounded and that the matrix $\sigma(t)$ has full rank for every t. Under the latter assumption the matrix $\sigma(\cdot)\sigma^*(\cdot)$ is invertible, so its inverse and the progressively measurable *relative risk process*

$$\vartheta(t)\triangleq\sigma^*(t)(\sigma(t)\sigma^*(t))^{-1}[b(t)-r(t)\mathbf{1}_m] \tag{2.4}$$

are well defined; here we denote by $\mathbf{1}_k$ the k-dimensional vector whose every component is one. We make the additional assumption that $\vartheta(\cdot)$ satisfies

$$E\int_0^T\|\vartheta(t)\|^2dt<\infty. \tag{2.5}$$

We shall encounter quite often the exponential local martingale process

$$Z(t)\triangleq\exp\left\{-\int_0^t\vartheta^*(s)dW(s)-\frac{1}{2}\int_0^t\|\vartheta(s)\|^2ds\right\}; \tag{2.6}$$

the discount process

$$\beta(t)\triangleq\exp\left\{-\int_0^t r(s)ds\right\}; \tag{2.7}$$

their product, that is, the so-called *state-price density process*

$$H(t) \triangleq \beta(t)Z(t); \tag{2.8}$$

as well as the process

$$W_0(t) \triangleq W(t) + \int_0^t \vartheta(s)ds, \qquad 0 \le t \le T. \tag{2.9}$$

We envision an economic agent who starts with a given initial endowment $x > 0$ and whose actions cannot affect the market prices. At any time $t \in [0,T]$ the agent can decide both the proportion $\pi_i(t)$ of his wealth $X(t)$ to be invested in the ith stock $(1 \le i \le m)$ and his consumption rate $c(t) \ge 0$. These decisions cannot anticipate the future but must depend only on the currently available information $\mathcal{F}(t)$. The remaining amount $[1 - \sum_{i=1}^m \pi_i(t)]X(t)$ is invested in the money market. Here the investor is allowed both to sell stocks short and to borrow money at the money market interest rate $r(\cdot)$; that is, the $\pi_i(\cdot)$ above are not restricted to take values only in $[0,1]$, and their sum may exceed 1.

The resulting *portfolio strategy* $\pi = (\pi_1, \dots, \pi_m)^* : [0,T] \times \Omega \to \mathbb{R}^m$ and *consumption strategy* $c : [0,T] \times \Omega \to [0,\infty)$ are assumed to be $\mathbb{F}$-progressively measurable processes and to satisfy the integrability condition $\int_0^T \big(c(t) + \|\pi(t)\|^2\big)dt < \infty$ almost surely.

According to the model dynamics of (2.1) and (2.2), the *wealth process* $X(\cdot) \equiv X^{x,\pi,c}(\cdot)$, corresponding to the portfolio/consumption pair (π, c) and initial capital $x \in \mathbb{R}^+$, is the solution of the linear stochastic differential equation

$$dX(t) = \sum_{i=1}^m \pi_i(t)X(t)\left\{b_i(t)dt + \sum_{j=1}^m \sigma_{ij}(t)dW_j(t)\right\} \tag{2.10}$$

$$+ \left\{1 - \sum_{i=1}^m \pi_i(t)\right\} X(t)r(t)dt - c(t)dt \tag{2.11}$$

$$= [r(t)X(t) - c(t)]dt + X(t)\pi^*(t)\sigma(t)dW_0(t),$$

subject to the initial condition $X(0) = x > 0$. Equivalently, we have

$$\beta(t)X(t) + \int_0^t \beta(s)c(s)ds = x + \int_0^t \beta(s)X(s)\pi^*(s)\sigma(s)dW_0(s), \tag{2.12}$$

and, from Itô's lemma, applied to the product of $Z(\cdot)$ and $\beta(\cdot)X(\cdot)$, we obtain

$$H(t)X(t) + \int_0^t H(s)c(s)ds = x + \int_0^t H(s)X(s)[\sigma^*(s)\pi(s) - \vartheta(s)]^*dW(s). \tag{2.13}$$

A portfolio/consumption process pair (π, c) is called *admissible for the initial capital* $x \in \mathbb{R}^+$ if the agent's wealth remains nonnegative at all times, i.e., if

$$X(t) \ge 0 \quad \forall\, t \in [0,T], \tag{2.14}$$

almost surely. We shall denote the family of admissible pairs (π, c) by $\mathcal{A}(x)$.

For any $(\pi, c) \in \mathcal{A}(x)$, the left-hand side of (2.13) is a continuous and nonnegative local martingale, thus a supermartingale. Consequently,

$$E\left(\int_0^T H(s)c(s)ds\right) \leq x \qquad \forall\ (\pi, c) \in \mathcal{A}(x). \tag{2.15}$$

Let us denote by $\mathcal{B}(x)$ the set of consumption policies $c : [0,T] \times \Omega \to [0, \infty)$ which are progressively measurable and satisfy (2.15). We have just verified that $c(\cdot) \in \mathcal{B}(x)$ for all pairs $(\pi, c) \in \mathcal{A}(x)$. In a complete market, where the number of stocks available for trading matches exactly the dimension of the "driving" Brownian motion, the converse holds as well, in the sense that any consumption strategy $c(\cdot)$ satisfying (2.15) can be financed by some portfolio policy $\pi(\cdot)$. For this reason, (2.15) can be interpreted as a "budget constraint".

LEMMA 2.1. *Let the market model of* (2.1), (2.2) *be complete, namely,* $m = d$. *Then, for every consumption process* $c(\cdot) \in \mathcal{B}(x)$, *there exists a portfolio process* $\pi(\cdot)$ *such that* $(\pi, c) \in \mathcal{A}(x)$, *and the associated wealth process* $X(\cdot) \equiv X^{x,\pi,c}(\cdot)$ *is given by*

$$H(t)X(t) = x + E_t(D(t)) - E(D(0)), \quad t \in [0,T], \qquad \textit{where} \quad D(t) \triangleq \int_t^T H(s)c(s)ds.$$

Here and in what follows, $E_t[\cdot]$ denotes conditional expectation $E[\cdot|\mathcal{F}(t)]$ with respect to the probability measure P, given the σ-algebra $\mathcal{F}(t)$. For the proof of Lemma 2.1, see Karatzas and Shreve (1998), pp. 166–169.

3. Utility functions. A *utility function* is a jointly continuous mapping $u : [0,T] \times \mathbb{R}^+ \to \mathbb{R}$ such that, for every $t \in [0,T]$, the function $u(t, \cdot)$ is strictly increasing, strictly concave, and of class $C^1(\mathbb{R}^+)$, and its derivative $u'(t,x) \triangleq \frac{\partial}{\partial x}u(t,x)$ satisfies

$$u'(t, 0^+) = \infty, \qquad u'(t, \infty) = 0. \tag{3.1}$$

These assumptions imply that the inverse $I(t, \cdot) : \mathbb{R}^+ \to \mathbb{R}^+$ of the function $u'(t, \cdot)$ exists for every $t \in [0,T]$, and is continuous and strictly decreasing with

$$I(t, 0^+) = \infty, \qquad I(t, \infty) = 0. \tag{3.2}$$

Furthermore, one can easily see the stronger assertion

$$\lim_{x\to\infty}\left(\max_{t\in[0,T]} u'(t,x)\right) = 0. \tag{3.3}$$

Let us now introduce, for each $t \in [0,T]$, the Legendre–Fenchel transform $\widetilde{u}(t, \cdot) : \mathbb{R}^+ \to \mathbb{R}$ of the convex function $-u(t, -x)$, namely,

$$\widetilde{u}(t,y) \triangleq \max_{x>0}\left[u(t,x) - xy\right] = u\big(t, I(t,y)\big) - yI(t,y), \qquad 0 < y < \infty. \tag{3.4}$$

The function $\widetilde{u}(t, \cdot)$ is strictly decreasing, is strictly convex, and satisfies

$$\frac{\partial}{\partial y}\widetilde{u}(t,y) = -I(t,y), \qquad 0 < y < \infty. \tag{3.5}$$

We note here that $\tilde{u} : [0,T] \times \mathbb{R}^+ \to \mathbb{R}$ is jointly continuous as well.

4. The maximization problem. For given utility function $u : [0,T] \times \mathbb{R}^+ \to \mathbb{R}$ and initial capital $x > 0$, we shall consider von Neumann–Morgenstern preferences with expected utility

$$J(z;\pi,c) \equiv J(z;c) \triangleq E\left[\int_0^T u\big(t, c(t) - z(t;c)\big)\,dt\right], \tag{4.1}$$

corresponding to any given pair $(\pi, c) \in \mathcal{A}(x)$ and its associated index process $z(\cdot) \equiv z(\cdot;c)$, defined in (1.2) and (4.3). This process represents the "standard of living" of the decision maker, an index that captures past consumption behavior and conditions the current consumption felicity by developing "habits". Of course, in order to ensure that the above expectation exists and is finite, we shall take into account only consumption strategies $c(\cdot)$ that satisfy

$$c(t) - z(t;c) > 0 \quad \forall\ 0 \le t \le T \tag{4.2}$$

almost surely. This additional budget specification insists that consumption must always exceed the standard of living, establishing incentives for a systematic buildup of habits over time and leading to "addiction patterns".

We shall stipulate that the standard of living follows the dynamics

$$\begin{aligned} dz(t) &= \big(\delta(t)c(t) - \alpha(t)z(t)\big)dt, \qquad t \in [0,T], \\ z(0) &= z. \end{aligned} \tag{4.3}$$

Here $\alpha(\cdot)$ and $\delta(\cdot)$ are nonnegative, bounded, and $\mathbb{F}$-adapted processes and $z \ge 0$ is a given real number. Thus, there exist constants $A > 0$ and $\Delta > 0$ such that

$$0 \le \alpha(t) \le A, \quad 0 \le \delta(t) \le \Delta \qquad \forall\ t \in [0,T] \tag{4.4}$$

hold almost surely. Equivalently, (4.3) stipulates $z(\cdot) \equiv z(\cdot;c)$ as in (1.2) and expresses $z(\cdot)$ as an exponentially weighted average of past consumption.

It is not hard to see that the constraint (4.2) forces the consumption $c(\cdot)$ to always exceed the so-called subsistence consumption $\widehat{c}(\cdot)$, for which $\widehat{c}(\cdot) \equiv z(\cdot;\widehat{c})$. This is the consumption pattern that matches exactly the standard of living at all times; from (4.3), this subsistence consumption satisfies the linear equation

$$d\,\widehat{c}(t) = \big(\delta(t) - \alpha(t)\big)\widehat{c}(t)dt, \quad t \in [0,T], \qquad \text{and} \qquad \widehat{c}(0) = z,$$

and we claim that, with $\widehat{z}(\cdot) \triangleq z(\cdot;\widehat{c})$, the constraint (4.2) implies that

$$c(t) > \widehat{c}(t) = \widehat{z}(t) = z\, e^{\int_0^t (\delta(v) - \alpha(v))dv} \quad \forall\ t \in [0,T]. \tag{4.5}$$

Indeed, in light of (1.2) the inequality (4.2) can be cast as

$$Q(t) \triangleq B(t) - \int_0^t B(s)\delta(s)\,ds > \widehat{B}(t) - \int_0^t \widehat{B}(s)\delta(s)\,ds \triangleq \widehat{Q}(t), \qquad 0 \le t \le T,$$

where $B(t) \triangleq c(t)\, e^{\int_0^t \alpha(s)ds}$, $\widehat{B}(t) \triangleq \widehat{c}(t)\, e^{\int_0^t \alpha(s)ds}$. A straightforward integration by parts deduces from this inequality the comparison

$$B(t) = Q(t) + \int_0^t Q(s)\, e^{\int_s^t \delta(u)du}\, ds > \widehat{Q}(t) + \int_0^t \widehat{Q}(s)\, e^{\int_s^t \delta(u)du}\, ds = \widehat{B}(t),$$

namely, the inequality claimed in (4.5).

Back into the budget constraint (2.15), this inequality (4.5) gives $x > wz$, where

$$w \triangleq E\left[\int_0^T e^{\int_0^t (\delta(v)-\alpha(v))dv} H(t)dt\right] \tag{4.6}$$

represents the "marginal" cost of subsistence consumption per unit of standard of living. Therefore, we need to impose the following restriction on the initial capital x and the initial standard of living level z.

Assumption 4.1. In the notation of (4.6), the pair (x, z) belongs to the set

$$\mathcal{D} \triangleq \Big\{(\mathfrak{x}, \mathfrak{z}) \in \mathbb{R}^+ \times [0, \infty);\ \mathfrak{x} > w\mathfrak{z}\Big\}.$$

Definition 4.2. *The* dynamic optimization *problem is to maximize the expression of* (4.1) *over the class* $\mathcal{A}'(x, z)$ *of* admissible portfolio/consumption pairs $(\pi, c) \in \mathcal{A}(x)$ *that satisfy* (4.2) *and*

$$E\left[\int_0^T u^-\big(t, c(t) - z(t; c)\big)\, dt\right] < \infty \tag{4.7}$$

(here and in what follows, b^- denotes the negative part of the real number b). The value function of this problem will be denoted by

$$V(x, z) \triangleq \sup_{(\pi,c)\in\mathcal{A}'(x,z)} J(z; \pi, c), \qquad (x, z) \in \mathcal{D}. \tag{4.8}$$

Definition 4.3. *The* static optimization *problem is to maximize the expression* (4.1) *over the set* $\mathcal{B}'(x, z)$ *of* consumption processes $c(\cdot) \in \mathcal{B}(x)$ *that satisfy* (4.2) *and* (4.7). *The value function of this problem will be denoted by*

$$U(x, z) \triangleq \sup_{c(\cdot)\in\mathcal{B}'(x,z)} J(z; c), \qquad (x, z) \in \mathcal{D}. \tag{4.9}$$

We obtain from (2.15) that $V(x, z) \le U(x, z)$ for all $(x, z) \in \mathcal{D}$. In fact, equality prevails here: it suffices to solve only the static maximization problem, since for a *static* consumption optimizer process $c_0(\cdot) \in \mathcal{B}'(x, z)$ in (4.9) we can always construct, according to Lemma 2.1, a portfolio process $\pi_0(\cdot)$ such that $(\pi_0, c_0) \in \mathcal{A}'(x, z)$ satisfies

$$U(x, z) = J(z; c_0) = J(z; c_0, \pi_0) = V(x, z) \quad \forall\ (x, z) \in \mathcal{D}$$

and constitutes a *dynamic* portfolio/consumption maximizing process pair for (4.8).

We also note that the set $\mathcal{B}'(x, z)$ of Definition 4.3 is convex, thanks to the linearity of $c \mapsto z(t; c)$ and the concavity of $x \mapsto u(t, x)$.

5. Solution of the optimization problem in complete markets. The static optimization problem of Definition 4.3 is treated as a typical maximization problem with constraints (2.15) and (4.2) in the case $m = d$ of a complete market and admits a solution derived by Detemple and Zapatero (1992). In this section, we shall follow briefly their analysis, obtaining further results associated with the value function V and with related features of the problem. More precisely, we shall identify the *effective state space* for the vector of wealth/standard of living processes, generated by the optimal portfolio/consumption pair, as a *random wedge* spanned by the temporal variable $t \in [0, T]$ and a family of suitable random half-planes (cf. Theorem 5.5).

Theorem 5.8 describes the relation of the value function V with a utility function as defined in section 3 and begins the study of its dual value function $\widetilde{V}$. An alternative representation for the quantity w of (4.6) is provided as well.

In providing constructive arguments for the existence of an optimal consumption policy to the static problem, a prominent role will be played by the "adjusted" (for the formation of habits) state-price density process

$$\Gamma(t) \triangleq H(t) + \delta(t) \cdot E_t\left(\int_t^T e^{\int_t^s (\delta(v)-\alpha(v))dv} H(s)ds\right), \qquad t \in [0,T], \tag{5.1}$$

which solves the *recursive linear stochastic equation*

$$\Gamma(t) = H(t) + \delta(t) \cdot E_t\left(\int_t^T e^{-\int_t^s \alpha(v)dv} \Gamma(s)ds\right), \quad t \in [0,T]; \tag{5.2}$$

cf. Detemple and Zapatero (1992). The process $\Gamma(\cdot)$ is the state-price density process $H(\cdot)$ compensated by an additional term that reflects the effect of habits. (In the absence of habits, that is, with $\delta(\cdot) \equiv 0$, we have $\Gamma(\cdot) \equiv H(\cdot)$.) Furthermore, we shall need to impose the following condition.

Assumption 5.1. It will be assumed that, for every $y \in \mathbb{R}^+$, we have

$$E\left(\int_0^T H(t)I(t, y\Gamma(t))\,dt\right) < \infty \qquad \text{and} \qquad E\left(\int_0^T \left|u\big(t, I(t, y\Gamma(t))\big)\right| dt\right) < \infty.$$

In what follows we shall provide conditions, on both the utility preferences and the model coefficients, which ensure the validity of the above assumption; cf. Remarks 5.7 and 6.3. Under Assumption 5.1, the function

$$\mathcal{X}(y) \triangleq E\left[\int_0^T \Gamma(t)I(t, y\Gamma(t))dt\right], \qquad 0 < y < \infty, \tag{5.3}$$

inherits from $I(t,\cdot)$ its continuity and strict decrease, as well as $\mathcal{X}(0^+) = \infty$ and $\mathcal{X}(\infty) = 0$. We shall denote the (continuous, strictly decreasing, onto) inverse of this function by $\mathcal{Y}(\cdot)$. Obviously, then Assumption 4.1 ensures the existence of a number $y_0 \triangleq \mathcal{Y}(x - wz) \in \mathbb{R}^+$ that satisfies

$$\mathcal{X}(y_0) = x - wz. \tag{5.4}$$

With this $y_0 > 0$, we consider now the process of net consumption given by

$$c_0(t) - z(t; c_0) \triangleq I\big(t, y_0\Gamma(t)\big) \quad \text{for} \;\; t \in [0,T]. \tag{5.5}$$

Inverting (5.5), we derive the relationship

$$\Gamma(t) = \frac{1}{y_0}\, u'\big(t, c_0(t) - z(t; c_0)\big), \qquad t \in [0,T], \tag{5.6}$$

which identifies the "adjusted" state-price density process $\Gamma(\cdot)$ as a "normalized marginal utility" process. Substituting back into (4.3), the standard of living process $z_0(\cdot) \equiv z(\cdot; c_0)$ is seen to satisfy the dynamics

$$dz_0(t) = \big[\delta(t)I(t, y_0\Gamma(t)) + (\delta(t) - \alpha(t))z_0(t)\big]dt, \qquad z_0(0) = z; \tag{5.7}$$

whereas solving the first-order linear ordinary differential equation (5.7) we arrive at the expression

$$z_0(t) = e^{\int_0^t (\delta(v)-\alpha(v))dv} \left[z + \int_0^t \delta(s) F_0(s) ds \right], \quad t \in [0,T], \tag{5.8}$$

with

$$F_0(t) \triangleq e^{\int_0^t (\alpha(v)-\delta(v))dv} I\big(t, y_0 \Gamma(t)\big), \quad t \in [0,T]. \tag{5.9}$$

Thanks to (5.5) and (5.8), we obtain the consumption process as

$$c_0(t) = e^{\int_0^t (\delta(v)-\alpha(v))dv} \left[F_0(t) + z + \int_0^t \delta(s) F_0(s) ds \right], \quad t \in [0,T]. \tag{5.10}$$

THEOREM 5.2. *Given Assumptions* 4.1 *and* 5.1, *the consumption process* $c_0(\cdot)$ *of* (5.10) *solves the static optimization problem of Definition* 4.3 *and satisfies the budget constraint* (2.15) *without slackness; that is,* $c_0(\cdot) \in \mathcal{B}'(x,z)$ *with*

$$E\left[\int_0^T H(t) c_0(t) dt \right] = x. \tag{5.11}$$

Furthermore, we have that $J(z;c) \le J(z;c_0) < \infty$ *holds for any* $c(\cdot) \in \mathcal{B}'(x,z)$.

Proof. From Assumption 5.1 and the identity (5.5), we see that $c_0(\cdot)$ satisfies condition (4.7) as well as $J(z;c_0) < \infty$. On the other hand, using (5.10) we obtain

$$\begin{aligned} E\left[\int_0^T H(t) c_0(t) dt \right] &= E\left[\int_0^T e^{\int_0^t (\delta(v)-\alpha(v))dv} H(t) \left(F_0(t) + z \right) dt \right. \\ &\qquad \left. + \int_0^T e^{\int_0^t (\delta(v)-\alpha(v))dv} H(t) \left(\int_0^t \delta(s) F_0(s) ds \right) dt \right] \\ &= E\left[\int_0^T e^{\int_0^t (\delta(v)-\alpha(v))dv} H(t) \left(F_0(t) + z \right) dt \right. \\ &\qquad \left. + \int_0^T \delta(t) F_0(t) \cdot E_t \left(\int_t^T e^{\int_0^s (\delta(v)-\alpha(v))dv} H(s) ds \right) dt \right] \\ &= E\left[\int_0^T \Gamma(t) I(t, y_0 \Gamma(t)) dt \right] + wz = x. \end{aligned}$$

(The next-to-last equation comes from the definitions (5.1), (5.9), whereas the last equation is a consequence of (5.3), (5.4).) The property (4.2) is also satisfied by $c_0(\cdot)$, thanks to (5.5) and to the property of infinite marginal utility at the origin, imposed in (3.1). It follows readily that $c_0(\cdot) \in \mathcal{B}'(x,z)$.

A proof for the last assertion of the theorem was given by Detemple and Karatzas (2003), in the case of nonaddictive habits. □

Remark 5.3. From (5.4), (5.11), (5.3), (5.5), (5.2), and (1.2), we have for $z > 0$ the computations

$$zw = E \int_0^T H(t) c_0(t) dt - \mathcal{X}(y_0) = E \int_0^T \Big[H(t) c_0(t) - \Gamma(t) \big(c_0(t) - z_0(t) \big) \Big] dt$$

$$= E\int_0^T \left[-\delta(t)\, E_t\left(\int_t^T e^{-\int_t^s \alpha(v)dv}\Gamma(s)ds\right)c_0(t) + z_0(t)\Gamma(t)\right] dt$$

$$= E\left[-\int_0^T \Gamma(s)\left(\int_0^s \delta(t)e^{-\int_t^s \alpha(v)dv}c_0(t)dt\right)ds + \int_0^T z_0(t)\Gamma(t)dt\right]$$

$$= E\int_0^T \left(z_0(t) - \int_0^t \delta(s)e^{-\int_s^t \alpha(v)dv}c_0(s)ds\right)\Gamma(t)\, dt$$

$$= z\cdot E\int_0^T e^{-\int_0^t \alpha(v)dv}\,\Gamma(t)\, dt.$$

We obtain the expression

$$w = E\left[\int_0^T e^{-\int_0^t \alpha(v)dv}\,\Gamma(t)\, dt\right], \tag{5.12}$$

which recasts the "subsistence-consumption-cost-per-unit-of-standard-of-living" quantity w of (4.6) as a weighted average of the "adjusted state-price density process" $\Gamma(\cdot)$ of (5.1), discounted at the rate $\alpha(\cdot)$.

This representation (5.12) of w makes the terminology "adjusted state-price density" for $\Gamma(\cdot)$ quite intuitive: namely, a comparison of (5.12) with (4.6), which involves only the density process $H(\cdot)$, suggests the significance of $\Gamma(\cdot)$ as a modified state-price density process that takes habit formation into account.

Corollary 5.4. *Under Assumptions* 4.1 *and* 5.1, *there exists a portfolio process* $\pi_0(\cdot)$ *such that the pair* $(\pi_0, c_0) \in \mathcal{A}'(x,z)$ *attains the supremum of* $J(z;\pi,c)$ *over* $\mathcal{A}'(x,z)$ *in* (4.8), *and the corresponding wealth process* $X_0(\cdot) \equiv X^{x,\pi_0,c_0}(\cdot)$ *is given by*

$$X_0(t) = \frac{1}{H(t)}\, E_t\left[\int_t^T H(s)c_0(s)\, ds\right], \quad t\in[0,T]. \tag{5.13}$$

This optimal investment $\pi_0(\cdot)$ *has the representation*

$$\pi_0(t) = \big(\sigma(t)\sigma^*(t)\big)^{-1}\sigma(t)\left[\frac{\psi_0(t)}{X_0(t)H(t)} + \vartheta(t)\right], \tag{5.14}$$

in terms of the $\mathbb{R}^d$*-valued,* $\mathbb{F}$*-progressively measurable, almost surely square-integrable process* $\psi_0(\cdot)$ *that represents the martingale*

$$M_0(t) \triangleq E_t\left[\int_0^T H(s)c_0(s)\, ds\right], \quad t\in[0,T], \tag{5.15}$$

as a stochastic integral, namely, $M_0(t) = x + \int_0^t \psi_0^*(s)dW(s)$. *Furthermore, the value function* V *of the dynamic maximization problem* (4.8) *is given as*

$$V(x,z) = G\big(\mathcal{Y}(x-wz)\big), \quad (x,z)\in\mathcal{D}; \tag{5.16}$$

here $\mathcal{Y}(\cdot)$ *is the inverse of the function* $\mathcal{X}(\cdot)$, *defined in* (5.3), *and*

$$G(y) \triangleq E\left[\int_0^T u\big(t, I(t, y\Gamma(t))\big)\, dt\right], \quad y\in\mathbb{R}^+. \tag{5.17}$$

Proof. The existence of the optimal portfolio $\pi_0(\cdot)$, along with the validation of (5.13)–(5.15), is a consequence of Lemma 2.1 and (5.11). From the optimality of (π_0, c_0) we get

$$V(x,z) = E\left[\int_0^T u\big(t, c_0(t) - z_0(t)\big)\,dt\right], \quad (x,z) \in \mathcal{D},$$

and (5.16) follows readily from (5.5), (5.4). □

The optimal policies (π_0, c_0) drive the investor to bankruptcy at time $t = T$: we have $X_0(T) = 0$ almost surely. This is natural, since utility is derived here only from consumption, not from terminal wealth.

Assumption 4.1 determines the "domain of acceptability" $\mathcal{D}$ for the initial values of wealth and standard of living. The next issue to be explored is the temporal evolution of these quantities as random processes, under the optimal pair policy (π_0, c_0) and for all times $t \in [0,T]$.

THEOREM 5.5. *Under Assumptions* 4.1 *and* 5.1*, the effective state space for the optimal wealth/standard of living process* $\big(X_0(\cdot), z_0(\cdot)\big)$ *is given by the family of* random wedges

$$\begin{aligned} \mathcal{D}_t &\triangleq \Big\{(\mathfrak{x},\mathfrak{z}) \in \mathbb{R}^+ \times [0,\infty);\ \mathfrak{x} > \mathcal{W}(t)\mathfrak{z}\Big\}, \qquad 0 \le t < T, \\ \mathcal{D}_T &\triangleq \Big\{(0,\mathfrak{z});\ \mathfrak{z} \in [0,\infty)\Big\}, \end{aligned} \tag{5.18}$$

where

$$\mathcal{W}(t) \triangleq \frac{1}{H(t)}\, E_t\left[\int_t^T e^{\int_t^s(\delta(v)-\alpha(v))dv} H(s)\,ds\right], \qquad 0 \le t \le T, \tag{5.19}$$

stands for the cost of subsistence consumption, per unit of standard of living, at time t*. In other words, we have, almost surely,*

$$\big(X_0(t), z_0(t)\big) \in \mathcal{D}_t \quad \forall\ t \in [0,T]. \tag{5.20}$$

Note that we have $\mathcal{W}(0) = w$ and $\mathcal{D}_0 = \mathcal{D}$, the quantities of Assumption 4.1; the random wedges $\mathcal{D}_t$ of (5.18) determine dynamically the range where the vector process of wealth/standard of living $\big(X_0(\cdot), z_0(\cdot)\big)$ takes values under the optimal regime.

Proof of Theorem 5.5. Consider the optimal pair (π_0, c_0) and the resulting standard of living process $z_0(\cdot)$, specified, respectively, by (5.14), (5.10), and (5.8). Recalling the definitions of (5.1) and (5.19), the corresponding wealth process $X_0(\cdot)$ of (5.13) may be reformulated as

$$\begin{aligned} X_0(t) = \frac{1}{H(t)}\, E_t\Bigg[\int_t^T H(s)\Bigg\{ & I(s, y_0\Gamma(s)) + z e^{\int_0^s(\delta(v)-\alpha(v))dv} \\ & + \int_0^s \delta(\theta) e^{\int_\theta^s(\delta(v)-\alpha(v))dv} I(\theta, y_0\Gamma(\theta))d\theta\Bigg\} ds\Bigg] \\ = \frac{1}{H(t)}\, E_t\Bigg[\int_t^T H(s)\Bigg\{ & z e^{\int_0^s(\delta(v)-\alpha(v))dv} \end{aligned}$$

$$
\begin{aligned}
&\qquad\qquad + \int_0^t \delta(\theta) e^{\int_\theta^s (\delta(v)-\alpha(v))dv} I(\theta, y_0\Gamma(\theta)) d\theta \Bigg\} ds \\
&\quad + \int_t^T H(s) I(s, y_0\Gamma(s)) ds \\
&\quad + \int_t^T \delta(\theta) I(\theta, y_0\Gamma(\theta)) \left(\int_\theta^T e^{\int_\theta^s (\delta(v)-\alpha(v))dv} H(s) ds \right) d\theta \Bigg] \\
&= \frac{1}{H(t)} \, E_t \Bigg[z_0(t) \int_t^T e^{\int_t^s (\delta(v)-\alpha(v))dv} H(s) ds \\
&\quad + \int_t^T \left\{ H(s) + \delta(s) E_s \left(\int_s^T H(\theta) e^{\int_s^\theta (\delta(v)-\alpha(v))dv} d\theta \right) \right\} I(s, y_0\Gamma(s)) ds \Bigg] \\
&= \mathcal{W}(t) z_0(t) + \frac{1}{H(t)} E_t \left[\int_t^T \Gamma(s) I(s, y_0\Gamma(s)) ds \right], \quad 0 \le t \le T.
\end{aligned}
$$

Therefore,

$$
X_0(t) - \mathcal{W}(t) z_0(t) = \frac{1}{H(t)} \, E_t \left[\int_t^T \Gamma(s) I(s, y_0\Gamma(s)) ds \right] > 0 \quad \forall\, t \in [0, T),
$$

almost surely, and (5.20) holds on $[0, T)$. The remaining assertions of the theorem follow directly from (5.13). □

Example 5.6 (*logarithmic utility*). Consider $u(t, x) = \log x$ for all $(t, x) \in [0, T] \times \mathbb{R}^+$. Then we have that $I(t, y) = 1/y$ for $(t, y) \in [0, T] \times \mathbb{R}^+$, $\mathcal{X}(y) = T/y$ for $y \in \mathbb{R}^+$, and $\mathcal{Y}(x) = T/x$ for $x \in \mathbb{R}^+$. The optimal consumption, standard of living, and wealth processes are as follows:

$$
\begin{aligned}
c_0(t) &= z\, e^{\int_0^t (\delta(v)-\alpha(v))dv} + \frac{x - wz}{T} \left[\frac{1}{\Gamma(t)} + \int_0^t \frac{\delta(s)}{\Gamma(s)} \, e^{-\int_s^t (\delta(v)-\alpha(v))dv} ds \right], \\
z_0(t) &= z\, e^{\int_0^t (\delta(v)-\alpha(v))dv} + \frac{x - wz}{T} \int_0^t \frac{\delta(s)}{\Gamma(s)} \, e^{-\int_s^t (\delta(v)-\alpha(v))dv} \, ds, \\
X_0(t) &= \frac{1}{H(t)} \left[z_0(t) E_t \left(\int_t^T e^{\int_t^s (\delta(v)-\alpha(v))dv} H(s) ds \right) + \frac{T - t}{T} (x - wz) \right]
\end{aligned}
$$

for $0 \le t \le T$. Moreover,

$$
G(y) = -T \log y - E \left[\int_0^T \log \Gamma(t) dt \right], \quad y \in \mathbb{R}^+,
$$

and the value function is

$$
V(x, z) = T \log \left(\frac{x - wz}{T} \right) - E \left[\int_0^T \log \Gamma(t) dt \right], \quad (x, z) \in \mathcal{D}.
$$

Note here that the conditions of Assumption 5.1 are satisfied; the first holds trivially, and the second is implied by the observation

$$E\big(\log \Gamma(t)\big) \le \log\big(E(\Gamma(t))\big) \le \varrho + \log\big(1 + \Delta T e^{\Delta T}\big) < \infty, \quad 0 \le t \le T,$$

where we used Jensen's inequality, (2.3), and the supermartingale property of $Z(\cdot)$. Finally, one may ascertain an explicit stochastic integral representation for $M_0(\cdot)$, defined in (5.15), under the additional assumption of *deterministic* model coefficients; cf. Example 7.9. The optimal portfolio process $\pi_0(\cdot)$ then follows by (5.14).

Remark 5.7. Consider utility functions such that

$$\sup_{0\le t\le T} I(t,y) \le \kappa y^{-\rho} \quad \forall\, y \in \mathbb{R}^+ \tag{5.21}$$

holds for some $\kappa > 0$, $\rho > 0$. Then the first condition of Assumption 5.1 holds under *at least one* of the subsequent conditions:

$$0 < \rho \le 1 \tag{5.22}$$

or

$$\vartheta(\cdot) \text{ is bounded uniformly on } [0,T]\times\Omega. \tag{5.23}$$

In particular, (5.21) and (5.22) yield

$$\mathcal{X}(y) \le \kappa y^{-\rho} E\left[\int_0^T (1 \vee \Gamma(t))\right] < \infty, \quad y \in \mathbb{R}^+.$$

Otherwise, use (5.23), (2.3), and the Novikov condition to set $(H(t))^{1-\rho} = m(t)L(t)$, in terms of the uniformly bounded process

$$m(t) \triangleq \exp\left\{(\rho-1)\int_0^t r(v)dv + \frac{1}{2}\rho(\rho-1)\int_0^t \|\vartheta(v)\|^2 dv\right\}$$

and the martingale

$$L(t) \triangleq \exp\left\{(\rho-1)\int_0^t \vartheta^*(v)dW(v) - \frac{1}{2}(\rho-1)^2\int_0^t \|\vartheta(v)\|^2 dv\right\}.$$

Then (5.21) implies that

$$\mathcal{X}(y) \le \kappa y^{-\rho}\left(1 + \Delta T e^{\varrho+\Delta T}\right)^{(1-\rho)} E\left[\int_0^T m(t)L(t)dt\right] < \infty, \quad y \in \mathbb{R}^+.$$

The function $V(\cdot, z)$ has all of the properties of a utility function as defined in section 3 for any given $z \ge 0$; we formalize this aspect of the value function in the result that follows, leading to the notion of a *generalized utility function* and to the explicit computation of its *convex dual*

$$\widetilde{V}(y) \triangleq \sup_{(x,z)\in\mathcal{D}} \Big\{V(x,z) - (x - wz)y\Big\}, \quad y \in \mathbb{R}. \tag{5.24}$$

THEOREM 5.8. *Under Assumptions* 4.1 *and* 5.1, *the mapping* $V : \mathcal{D} \to \mathbb{R}$ *of* (4.8) *is a generalized utility function, in the sense of being strictly concave and of class*

$C^{1,1}(\mathcal{D})$; it is strictly increasing in its first argument, is strictly decreasing in the second, and satisfies $V_x((wz)^+, z) = \infty$, $V_x(\infty, z) = 0$ for any $z \geq 0$. Furthermore,

$$\lim_{\substack{(x,z)\to(\chi,\zeta)\\(x,z)\in\mathcal{D}}} V(x,z) = \int_0^T u(t, 0^+)\,dt \qquad \forall\ (\chi,\zeta) \in \partial\mathcal{D}, \tag{5.25}$$

where $\partial\mathcal{D} = \{(\mathfrak{x},\mathfrak{z}) \in [0,\infty)^2;\ \mathfrak{x} = w\,\mathfrak{z}\}$ is the boundary of $\mathcal{D}$.

Furthermore, with $\mathcal{X}(\cdot)$ and $G(\cdot)$ given by (5.3) *and* (5.17)*, respectively, we have*

$$V_x(x,z) = \mathcal{Y}(x - wz), \qquad V_z(x,z) = -w\mathcal{Y}(x - wz) \quad \forall\ (x,z) \in \mathcal{D}, \tag{5.26}$$

$$\widetilde{V}(y) = G(y) - y\mathcal{X}(y) = E\int_0^T \widetilde{u}\big(t, y\Gamma(t)\big)\,dt \quad \forall\ y > 0, \tag{5.27}$$

$$\widetilde{V}'(y) = -\mathcal{X}(y) \quad \forall\ y > 0. \tag{5.28}$$

Proof. We show first the strict concavity of V. Let $(x_1, z_1), (x_2, z_2) \in \mathcal{D}$ and $\lambda_1, \lambda_2 \in (0,1)$ such that $\lambda_1 + \lambda_2 = 1$. For each (x_i, z_i) consider the optimal portfolio/consumption policy $(\pi_i, c_i) \in \mathcal{A}'(x_i, z_i)$ which generates the corresponding wealth process $X^{x_i,\pi_i,c_i}(\cdot)$ and the standard of living process $z_i(\cdot)$, $i = 1, 2$. Define now the portfolio/consumption plan $(\pi, c) \triangleq (\lambda_1\pi_1 + \lambda_2\pi_2, \lambda_1 c_1 + \lambda_2 c_2)$, denoting by $X^{x,\pi,c}(\cdot)$, $z(\cdot)$ the corresponding wealth and standard of living with $x \triangleq \lambda_1 x_1 + \lambda_2 x_2$ and $z \triangleq \lambda_1 z_1 + \lambda_2 z_2$. It is then easy to see that $(\pi, c) \in \mathcal{A}'(x,z)$ and

$$X^{x,\pi,c}(\cdot) = \lambda_1 X^{x_1,\pi_1,c_1}(\cdot) + \lambda_2 X^{x_2,\pi_2,c_2}(\cdot), \quad z(\cdot) = \lambda_1 z_1(\cdot) + \lambda_2 z_2(\cdot)$$

hold almost surely. Therefore, the strict concavity of $u(t,\cdot)$ implies that

$$\begin{aligned} &\lambda_1 V(x_1, z_1) + \lambda_2 V(x_2, z_2) \\ &= \lambda_1 E\left[\int_0^T u(t, c_1(t) - z_1(t))dt\right] + \lambda_2 E\left[\int_0^T u(t, c_2(t) - z_2(t))dt\right] \\ &< E\left[\int_0^T u(t, c(t) - z(t))dt\right] \leq V(x,z) = V(\lambda_1 x_1 + \lambda_2 x_2, \lambda_1 z_1 + \lambda_2 z_2). \end{aligned}$$

As a real-valued concave function on $\mathcal{D}$, V is continuous on its domain.

To establish (5.25), we consider pairs $(x,z) \in \mathcal{D}$ and observe from (5.16) that $\lim_{(x,z)\to(\chi,\zeta)} V(x,z) = \lim_{y\to\infty} G(y)$ holds for any $(\chi,\zeta) \in \partial\mathcal{D}$. But (3.2) indicates that $\lim_{y\to\infty} I(t, y\Gamma(t)) = 0$ for $0 \leq t \leq T$, and Assumption 5.1 ensures that $G(y)$ of (5.17) is finite for any $y \in \mathbb{R}^+$; thus, (5.25) becomes a direct consequence of the monotone convergence theorem.

We undertake (5.27) next; its second equality is checked algebraically via (3.4), (5.3), and (5.17). Turning now to the first, for every $(x,z) \in \mathcal{D}$, $y > 0$, and $(\pi, c) \in \mathcal{A}'(x,z)$, the relation of (3.4) gives

$$u\big(t, c(t) - z(t)\big) \leq \widetilde{u}(t, y\Gamma(t)) + y\Gamma(t)\big(c(t) - z(t)\big). \tag{5.29}$$

Taking expectations, we use (1.2), (5.2), (5.12), and the budget constraint (2.15) to obtain

$$
\begin{aligned}
E\int_0^T u\big(t,c(t)-z(t)\big)\,dt &\le E\int_0^T \Big[\widetilde{u}\big(t,y\Gamma(t)\big)+y\Gamma(t)\big(c(t)-z(t)\big)\Big]dt\\
&= E\int_0^T \widetilde{u}\big(t,y\Gamma(t)\big)dt\\
&\quad + y\cdot E\int_0^T \Gamma(t)\left(c(t)-ze^{-\int_0^t\alpha(v)dv}-\int_0^t\delta(s)e^{-\int_s^t\alpha(v)dv}c(s)ds\right)dt\\
&= E\int_0^T \widetilde{u}\big(t,y\Gamma(t)\big)dt-ywz\\
&\quad + y\cdot E\left[\int_0^T\Gamma(t)c(t)dt-\int_0^T\delta(s)\left(\int_s^T e^{-\int_s^t\alpha(v)dv}\Gamma(t)dt\right)c(s)ds\right]\qquad(5.30)\\
&= E\int_0^T \widetilde{u}\big(t,y\Gamma(t)\big)\,dt-ywz\\
&\quad + y\cdot E\int_0^T\left\{\Gamma(t)-\delta(t)E_t\left(\int_t^T e^{-\int_t^s\alpha(v)dv}\Gamma(s)ds\right)\right\}c(t)dt\\
&= E\int_0^T \widetilde{u}\big(t,y\Gamma(t)\big)\,dt-ywz+y\cdot E\int_0^T H(t)c(t)dt\\
&\le E\int_0^T \widetilde{u}\big(t,y\Gamma(t)\big)dt+y(x-wz)=G(y)-y\mathcal{X}(y)+y(x-wz).
\end{aligned}
$$

The inequalities in (5.30) will hold as equalities if and only if

$$
c(t)-z(t)=I\big(t,y\Gamma(t)\big)\quad\text{and}\quad E\int_0^T H(t)c(t)\,dt=x. \tag{5.31}
$$

Setting $Q(y)\triangleq G(y)-y\mathcal{X}(y)$ and maximizing over $(\pi,c)\in\mathcal{A}'(x,z)$, it follows from (5.30) that $V(x,z)\le Q(y)+(x-wz)y$ for every $(x,z)\in\mathcal{D}$, and thereby $\widetilde{V}(y)\le Q(y)$ for every $y>0$. Conversely, (5.30) becomes an equality if the first equation of (5.31) is satisfied and if $\mathcal{X}(y)=x-wz$, so $Q(y)=V(\mathcal{X}(y)+wz,z)-\mathcal{X}(y)y\le\widetilde{V}(y)$. Hence (5.27) is established, and clearly the supremum in (5.24) is attained if $x-wz=\mathcal{X}(y)$.

We argue now (5.28) by noting the identity

$$
\begin{aligned}
yI(t,y)-hI(t,h)-\int_h^y I(t,\lambda)d\lambda &= yI(t,y)-hI(t,h)+\widetilde{u}(t,y)-\widetilde{u}(t,h)\\
&= u(t,I(t,y))-u(t,I(t,h)),\qquad(5.32)
\end{aligned}
$$

which holds for any utility function $u(\cdot)$ and $0\le t\le T$, $0<h<y<\infty$; recall (3.4) and (3.5). This enables us to compute

$$y\mathcal{X}(y) - h\mathcal{X}(h) - \int_h^y \mathcal{X}(\xi)d\xi$$

$$(5.33) \qquad = E\int_0^T \left[yH(t)I(t,yH(t)) - hH(t)I(t,hH(t)) - \int_{hH(t)}^{yH(t)} I(t,\lambda)d\lambda \right] dt$$

$$= E\int_0^T \Big[u\big(t,I(t,yH(t))\big) - u\big(t,I(t,hH(t))\big)\Big] dt = G(y) - G(h),$$

which in conjunction with (5.27) leads to

$$(5.34) \qquad \widetilde{V}(y) - \widetilde{V}(h) = -\int_h^y \mathcal{X}(\xi)d\xi, \qquad 0 < h < y < \infty,$$

and (5.28) follows. Finally, let us rewrite (5.24) in the more suggestive form

$$\widetilde{V}(y) = \sup_{(x,z)\in\mathcal{D}} \Big\{ V(x,z) - (x,z)\cdot(y,-wy) \Big\}, \quad y \in \mathbb{R},$$

where $v_1 \cdot v_2$ stands for the dot product between any two vectors v_1 and v_2. We recall that for $(x^*,z^*) \in \mathcal{D}$ and $y > 0$ we have $(y,-wy) \in \partial V(x^*,z^*)$ if and only if the maximum in the above expression is attained by (x^*,z^*) (see, e.g., Rockafellar (1970), Theorem 23.5). However, we have already shown that this maximum is attained by the pair (x^*,z^*) only if $x^* - wz^* = \mathcal{X}(y)$, implying that

$$\partial V(x^*,z^*) = \Big\{ \big(\mathcal{Y}(x^* - wz^*), -w\mathcal{Y}(x^* - wz^*)\big) \Big\}.$$

This proves (5.26) (see, e.g., Theorem 23.4 of Rockafellar (1970)), and it implies that $V_x(\cdot,z)$ is continuous, positive (thus $V(\cdot,z)$ is strictly increasing), and strictly decreasing on (wz,∞), with $\lim_{x\downarrow wz} V_x(x,z) = \lim_{x\downarrow wz} \mathcal{Y}(x-wz) = \infty$ and $\lim_{x\uparrow\infty} V_x(x,z) = \lim_{x\uparrow\infty} \mathcal{Y}(x-wz) = 0$; meanwhile, $V_z(x,\cdot) < 0$ is continuous, so $V(x,\cdot)$ is strictly decreasing. Consequently, $V(\cdot,\cdot)$ is a generalized utility function. ◻

Remark 5.9. We note that, given any $z \in [0,\infty)$, (5.16) can be written as $G(y) = V(\mathcal{X}(y) + wz, z)$ for every $y \in \mathbb{R}^+$. Thus, if $\mathcal{X}(\cdot)$ is differentiable, then by (5.26) the function $G(\cdot)$ is also differentiable with

$$(5.35) \qquad G'(y) = V_x\big(\mathcal{X}(y) + wz, z\big)\mathcal{X}'(y) = y\,\mathcal{X}'(y), \quad y \in \mathbb{R}^+.$$

6. The role of SPDEs. In section 5 we established the existence and uniqueness, up to almost-everywhere equivalence, of a solution to our habit-modulated utility maximization problem in the case of a complete security market. The analysis provided a concrete representation for the optimal consumption process $c_0(\cdot)$, given by (5.10), but not for the optimal portfolio strategy $\pi_0(\cdot)$; no useful expression aside from (5.14) was given for it. In this section we shall address this issue using a technique based on the ideas of *dynamic programming.* Our motivation goes back to Theorem 5.5, which brings forth the *dynamic* nature of the optimal wealth/standard of living pair $\big(X_0(\cdot), z_0(\cdot)\big)$ in terms of a stochastically evolving range.

Our analysis will be supported by the theory of *backward stochastic partial differential equations* (BSPDEs) and their interrelation with appropriate *adapted versions of stochastic Feynman–Kac formulae*, which have been developed fairly recently. This interplay will be based on the generalized Itô–Kunita–Wentzell formula (cf. Proposition 8.1) and will show that the value function of problem (4.8) satisfies a nonlinear, backward *stochastic* Hamilton–Jacobi–Bellman partial differential equation of parabolic type.

We shall provide the optimal portfolio $\pi_0(t)$ and consumption policy $c_0(t)$ in closed, *stochastic* "feedback forms" on the current wealth $X_0(t)$ and the standard of living $z_0(t)$. In other words, we shall get hold of suitable *random fields* $C : [0,T) \times \mathbb{R}^+ \times [0,\infty) \times \Omega \to \mathbb{R}^+$ and $\Pi : [0,T) \times \mathbb{R}^+ \times [0,\infty) \times \Omega \to \mathbb{R}^d$, for which

$$(6.1) \qquad c_0(t) = C(t, X_0(t), z_0(t)) \quad \text{and} \quad \pi_0(t) = \Pi(t, X_0(t), z_0(t)), \qquad 0 \le t < T.$$

The conditions listed below will allow us to present the main concepts of our dynamic approach with a minimum of technical fuss.

Assumption 6.1. The model-coefficient processes $r(\cdot)$, $b(\cdot)$, $\vartheta(\cdot)$, $\sigma(\cdot)$, $\alpha(\cdot)$, and $\delta(\cdot)$ are continuous, and $\|\vartheta(\cdot)\|$ is bounded away from zero and infinity:

$$(6.2) \qquad \exists\ k_1, k_2 > 0 \ \text{ such that } \quad 0 < k_1 \le \|\vartheta(t)\| \le k_2 < \infty \ \ \forall\, t \in [0,T].$$

It will also be assumed that $\delta(\cdot)$ is differentiable and $r(\cdot) - \delta(\cdot) + \alpha(\cdot)$ is nonrandom.

These last assumptions on $\delta(\cdot)$ and $r(\cdot) - \delta(\cdot) + \alpha(\cdot)$ are rather severe and can actually be relaxed and omitted, respectively; cf. Remark 6.4. They will be crucial, however, in our effort to keep the required analysis and notation at manageable levels, without obscuring by technicalities the essential ideas.

Since the market price of risk $\vartheta(\cdot)$ is now assumed to be bounded, the local martingale $Z(\cdot)$ of (2.6) becomes a martingale. Thus, by Girsanov's theorem, the process $W_0(\cdot)$ of (2.9) is a standard, d-dimensional Brownian motion under the new probability measure

$$(6.3) \qquad P^0(A) \triangleq E[Z(T)\mathbf{1}_A], \quad A \in \mathcal{F}(T).$$

We shall refer to P^0 as the *equivalent martingale measure* of the financial market $\mathcal{M}_0$ and denote expectation under this measure by E^0.

Assumption 6.2. We shall assume that the utility function $u(\cdot)$ satisfies the following:

(i) polynomial growth of I:

$$\exists\ \gamma > 0 \quad \text{such that} \quad I(t,y) \le \gamma + y^{-\gamma} \quad \forall\ (t,y) \in [0,T] \times \mathbb{R}^+;$$

(ii) polynomial growth of u ∘ I:

$$\exists\ \gamma > 0 \quad \text{such that} \quad u\big(t, I(t,y)\big) \ge -\gamma - y^{\gamma} \ \ \forall\ (t,y) \in [0,T] \times \mathbb{R}^+;$$

(iii) for each $t \in [0,T]$, $y \mapsto u(t,y)$ and $y \mapsto I(t,y)$ are of class $C^4(\mathbb{R}^+)$;

(iv) $I'(t,y) = \frac{\partial}{\partial y} I(t,y)$ is strictly negative for every $(t,y) \in [0,T] \times \mathbb{R}^+$;

(v) for every $t \in [0,T]$, $y \mapsto g(t,y) \triangleq y I'(t,y)$ is increasing and concave.

Remark 6.3. Assumption 6.2(i),(ii), together with (3.3) and the strict decrease of $I(t,\cdot)$, yields that

$$\exists\ \gamma > 0 \quad \text{such that} \ \ \big|u\big(t, I(t,y)\big)\big| \le \gamma + y^{\gamma} + y^{-\gamma} \quad \forall\ (t,y) \in [0,T] \times \mathbb{R}^+.$$

Notice that Assumptions 6.1 and 6.2(i),(ii) guarantee the validity of Assumption 5.1 in the preceding section; compare also with Remark 5.7. Moreover, the composite function $u(t,I(t,\cdot))$ inherits the order of smoothness posited in Assumption 6.2(iii) for its components for every $t \in [0,T]$.

For each $(t,y) \in [0,T] \times \mathbb{R}^+$ and $t \le s \le T$, we consider the stochastic processes

$$Z^t(s) \triangleq \exp\left\{ - \int_t^s \vartheta^*(v)dW(v) - \frac{1}{2}\int_t^s \|\vartheta(v)\|^2 dv \right\}, \quad H^t(s) \triangleq Z^t(s)\, e^{-\int_t^s r(v)dv}. \tag{6.4}$$

These extend the processes of (2.6) and (2.8), respectively, to initial times other than zero. In accordance with (5.1), we shall also consider the extended "adjusted" state-price density process

$$\begin{aligned} \Gamma^t(s) &\triangleq H^t(s) + \delta(s) \cdot E_s \left(\int_s^T e^{\int_s^\theta (\delta(v)-\alpha(v))dv} H^t(\theta) d\theta \right) \\ &= H^t(s)\Big[1 + \delta(s)\mathcal{W}(s)\Big] \qquad (6.5) \\ &= H^t(s)\left[1 + \delta(s)\int_s^T e^{\int_s^\theta (-r(v)+\delta(v)-\alpha(v))dv} d\theta \right] = H^t(s)\mu(s), \quad t \le s \le T. \end{aligned}$$

We have used (5.19), Assumption 6.1, and the martingale property of $Z(\cdot)$ and have set

$$\mu(t) \triangleq 1 + \delta(t)w(t), \qquad \text{where} \qquad w(t) \triangleq \int_t^T e^{\int_t^s (-r(v)+\delta(v)-\alpha(v))dv} ds, \tag{6.6}$$

$$w'(t) \triangleq \frac{d}{dt}w(t) = \big[r(t) + \alpha(t) - \delta(t)\big]w(t) - 1 = \big[r(t) + \alpha(t)\big]w(t) - \mu(t) \tag{6.7}$$

for $0 \le t \le T$. Note that $w(\cdot)$ is the deterministic reduction of $\mathcal{W}(\cdot)$ in (5.19) under the simplifying assumption (Assumption 6.1); namely, $\mathcal{W}(\cdot) \equiv w(\cdot)$ within the context of this section.

We shall introduce also the diffusion process

$$Y^{(t,y)}(s) \triangleq y\Gamma^t(s), \qquad t \le s \le T, \tag{6.8}$$

which, from (6.4) and (6.5), satisfies the linear stochastic differential equation

$$dY^{(t,y)}(s) = Y^{(t,y)}(s)\left[\left(\frac{\mu'(s)}{\mu(s)} - r(s) \right) ds - \vartheta^*(s)dW(s) \right], \tag{6.9}$$

or equivalently

$$dY^{(t,y)}(s) = Y^{(t,y)}(s)\left[\left(\frac{\mu'(s)}{\mu(s)} - r(s) + \|\vartheta(s)\|^2 \right) ds - \vartheta^*(s)dW_0(s) \right] \tag{6.10}$$

with initial condition

$$Y^{(t,y)}(t) = y\mu(t), \tag{6.11}$$

as well as $Y^{(t,y)}(s) = yY^{(t,1)}(s) = yH(s)\mu(s)/H(t)$.

Remark 6.4. The last two conditions of Assumption 6.1 on $\delta(\cdot)$ and $r(\cdot) - \delta(\cdot) + \alpha(\cdot)$ allowed us to derive the useful representation (6.5) for the process $\Gamma^t(\cdot)$ with a minimum of notation and technical fuss. This representation led to the very explicit computation of the semimartingale decomposition for the process $Y^{(t,y)}(\cdot)$ in (6.8)–(6.11), which will be crucial for our analysis in the rest of the paper.

These two conditions will not be needed any further; in fact, it is possible actually to relax the former and omit the latter. In particular, we may instead assume that $\delta(\cdot)$ is a semimartingale and still obtain a representation of the form (6.5) and a semimartingale decomposition for the process $Y^{(t,y)}(\cdot)$ of (6.8).

To this end, it suffices to show that $\mathcal{W}(\cdot)$ of (5.19) has a semimartingale decomposition as well. Indeed,

$$\mathcal{W}(t) = \frac{1}{\widehat{H}(t)} E_t\left[\int_t^T \widehat{H}(s)ds\right] = \frac{1}{\widehat{H}(t)}\left\{E_t\left[\int_0^T \widehat{H}(s)ds\right] - \int_0^t \widehat{H}(s)ds\right\},$$

where we have $\widehat{H}(t) \triangleq e^{\int_0^t(\delta(v)-\alpha(v))dv}H(t)$ for $0 \le t \le T$. Since the process $\widehat{M}(t) \triangleq E_t \int_0^T \widehat{H}(s)ds$, $0 \le t \le T$, is a local martingale under the measure P, the standard representation result for Brownian local martingales as stochastic integrals (see, e.g., Karatzas and Shreve (1991), Problem 3.4.16) implies the existence of an $\mathbb{R}^d$-valued, $\mathbb{F}$-progressively measurable and almost sure square-integrable process $\widehat{\psi}(\cdot)$ so that $\widehat{M}(t) = \widehat{M}(0) + \int_0^t \widehat{\psi}(s)dW(s)$, $0 \le t \le T$.

This leads to a semimartingale decomposition for $\mathcal{W}(\cdot)$ and thence to a similar decomposition for the process $Y^{(t,y)}(\cdot) = yZ^t(\cdot)e^{-\int_t^\cdot r(v)dv}[1+\delta(\cdot)\mathcal{W}(\cdot)]$ in (6.8); this decomposition is far more cumbersome to write down, or keep track of, than (6.9); yet it exists. We have opted for the simplicity of (6.9), afforded by the conditions of Assumption 6.1 on $\delta(\cdot)$ and $r(\cdot) - \delta(\cdot) + \alpha(\cdot)$. □

Invoking the "Bayes rule" for conditional expectations, a computation similar to the one presented in the proof of Theorem 5.5 shows that the optimal wealth/standard of living vector process $\big(X_0(\cdot), z_0(\cdot)\big)$ of (5.8), (5.13) satisfies

$$\begin{aligned} X_0(t) - w(t)z_0(t) &= \frac{1}{\xi}\, E_t\Big[\int_t^T Y^{(t,\xi)}(s)\, I\big(s, Y^{(0,\xi)}(s)\big)\, ds\Big] \\ &= E_t^0\Big[\int_t^T e^{-\int_t^s r(v)dv}\mu(s)\, I\big(s, Y^{(0,\xi)}(s)\big)\, ds\Big] = \mathfrak{X}\left(t, \frac{Y^{(0,\xi)}(t)}{\mu(t)}\right) \end{aligned} \tag{6.12}$$

for $0 \le t \le T$ and $\xi = \mathcal{Y}(x - wz)$. We have used here the definition (6.8) and introduced the random field $\mathfrak{X} : [0,T] \times \mathbb{R}^+ \times \Omega \to \mathbb{R}^+$ defined as

$$\mathfrak{X}(t,y) \triangleq E_t^0\left[\int_t^T e^{-\int_t^s r(v)dv}\mu(s)\, I\big(s, yY^{(t,1)}(s)\big)\, ds\right]. \tag{6.13}$$

A comparison of (5.3), (6.12), and (6.13) reveals the dynamic and stochastic evolution of the function $\mathcal{X}(\cdot)$ as a random field, since $\mathcal{X}(\cdot) = \mathfrak{X}(0,\cdot)$.

We proceed with the derivation of the random fields C and Π in (6.1) by obtaining first a semimartingale decomposition for the random field $\mathfrak{X}(\cdot,\cdot)$ of (6.13). A significant role in this program will be played by a BSPDE established for $\mathfrak{X}(\cdot,\cdot)$, which will lead to a stochastic Feynman–Kac formula and consequently to the desired decomposition.

Remark 6.5. Similar results have been obtained by Ma and Yong (1997) in related contexts (integral representations, stochastic Black–Scholes formulae) using the so-called four-step scheme for forward-backward stochastic differential equations (FBSDEs). In the appendix, we formalize this connection by providing both a stochastic

Feynman–Kac result and an equivalent FBSDE formulation for the various BSPDEs constructed in the present and the next section.

LEMMA 6.6. *Consider the random field* $\mathfrak{X}$ *of* (6.13). *Under Assumptions* 6.1 *and* 6.2, *there exists a random field* $\Psi^{\mathfrak{X}} : [0,T] \times \mathbb{R}^+ \times \Omega \to \mathbb{R}^d$ *such that the pair* $(\mathfrak{X}, \Psi^{\mathfrak{X}})$ *belongs to the class* $C_{\mathbb{F}}\big([0,T]; \mathbb{L}^2(\Omega; C^3(\mathbb{R}^+))\big) \times \mathbb{L}^2_{\mathbb{F}}\big(0,T; \mathbb{L}^2(\Omega; C^2(\mathbb{R}^+;\mathbb{R}^d))\big)$ *and is the unique solution of the Cauchy problem for the equation*

$$\begin{aligned} -d\mathfrak{X}(t,y) = \Bigg[& \frac{1}{2}\|\vartheta(t)\|^2 y^2 \mathfrak{X}_{yy}(t,y) + \big(\|\vartheta(t)\|^2 - r(t)\big)\, y\mathfrak{X}_y(t,y) - r(t)\mathfrak{X}(t,y) \\ & - \vartheta^*(t) y \Psi^{\mathfrak{X}}_y(t,y) + \mu(t) I\big(t, y\mu(t)\big)\Bigg] dt - \big(\Psi^{\mathfrak{X}}(t,y)\big)^* \, dW_0(t) \end{aligned} \tag{6.14}$$

on $[0,T) \times \mathbb{R}^+$, *subject to the terminal condition*

$$\mathfrak{X}(T,y) = 0 \quad on \quad \mathbb{R}^+, \tag{6.15}$$

almost surely. Furthermore, for each $t \in [0,T)$ *we have* $\mathfrak{X}(t,0^+) = \infty$, $\mathfrak{X}(t,\infty) = 0$, *and* $\mathfrak{X}(t,\cdot)$ *is strictly decreasing; this establishes the existence of a strictly decreasing inverse random field* $\mathfrak{Y}(t,\cdot) : \mathbb{R}^+ \xrightarrow{onto} \mathbb{R}^+$, *such that*

$$\mathfrak{X}\big(t, \mathfrak{Y}(t,x)\big) = x \qquad \forall\, x \in \mathbb{R}^+ \tag{6.16}$$

holds almost surely. The resulting random field $\mathfrak{Y}$ *is of class* $C_{\mathbb{F}}\big([0,T); C^3(\mathbb{R}^+)\big)$.

Proof. The first part of the lemma is verified directly from (6.10), (6.11), (6.13), and Proposition 8.3 in the appendix, through the identifications

$$m(\cdot) = \frac{\mu'(\cdot)}{\mu(\cdot)} - r(\cdot) + \|\vartheta(\cdot)\|^2, \qquad \nu(\cdot) = -\vartheta(\cdot), \qquad N(\cdot,\cdot) = I(\cdot,\cdot),$$

$$\ell(\cdot) = q(\cdot) = \mu(\cdot), \qquad \text{and} \qquad \rho(\cdot) = -r(\cdot).$$

Next, we shall verify that $\mathfrak{X}_y(t,y)$ is strictly negative, almost surely. To this end, let $(t,y) \in [0,T) \times \mathbb{R}^+$, $h > 0$, and invoke the (strict) decrease of $I(t,\cdot)$, coupled with (2.3), to verify that

$$\frac{1}{h}\big[\mathfrak{X}(t,y) - \mathfrak{X}(t,y+h)\big] \geq E^0_t \left[\int_t^T \frac{e^{-\varrho}}{h} \Big\{ I\big(s, yY^{(t,1)}(s)\big) - I\big(s, (y+h)Y^{(t,1)}(s)\big) \Big\} ds \right].$$

By the mean value theorem, there is a real number $y_h \in [y, y+h]$ such that

$$I\big(s, yY^{(t,1)}(s)\big) - I\big(s, (y+h)Y^{(t,1)}(s)\big) = -hY^{(t,1)}(s) I'\big(s, y_h Y^{(t,1)}(s)\big),$$

and conditions (2.3), (4.4), (6.2), and the supermartingale property of $Z^t(\cdot)$ imply the inequality $Y^{(t,1)}(s) \leq \phi(s) Z^t_0(s)$, in terms of the deterministic function $\phi(t) \triangleq \big[1 + \Delta(T-s)\, e^{\varrho + \Delta(T-s)}\big] e^{\varrho + \kappa_2^2 (T-t)}$ and the P^0-martingale

$$Z^t_0(s) \triangleq \exp\left\{ - \int_t^s \vartheta^*(v) dW_0(v) - \frac{1}{2} \int_t^s \|\vartheta(v)\|^2 dv \right\}, \qquad t \leq s \leq T. \tag{6.17}$$

Due to Assumption 6.2(v), the right-hand side of the former inequality attains the lower bounds

$$-E_t^0\left[\int_t^T \frac{e^{-\varrho}}{y_h}\, g\big(s, y_h\phi(s)Z_0^t(s)\big)ds\right] \geq -\int_t^T \frac{e^{-\varrho}}{y_h}\, g\Big(s, y_h\phi(s)E_t^0\big(Z_0^t(s)\big)\Big)ds$$

$$= -e^{-\varrho}\int_t^T I'(s, y_h\phi(s))\phi(s)ds,$$

where we have also used Jensen's inequality. Passing to the limit as $h \downarrow 0$, we obtain from Fatou's lemma

$$\mathfrak{X}_y(t,y) \leq e^{-\varrho}\int_t^T I'(s, y\phi(s))\phi(s)ds < 0.$$

According to the implicit function theorem, the inverse $\mathfrak{Y}(t,\cdot) : \mathbb{R}^+ \xrightarrow{onto} \mathbb{R}^+$ of the random field $\mathfrak{X}(t,\cdot)$ exists almost surely, in the context of (6.16); in fact, the two random fields have the same order of regularity on their respective domains. Concluding, the claimed values of $\mathfrak{X}(t,0^+)$ and $\mathfrak{X}(t,\infty)$ are easily confirmed, respectively, by the monotone and dominated convergence theorem. □

Remark 6.7. It is worth noting that Lemma 6.6 assigns to the pair of random fields $(\mathfrak{X}, \Psi^{\mathfrak{X}})$ an additional order of smoothness than is required for solving the SPDE (6.14), (6.15). Nevertheless, this extra smoothness allows us to apply the Itô–Kunita–Wentzell formula of Proposition 8.1 in the appendix. Furthermore, the above lemma yields the representation

$$\mathfrak{X}(t,y) = \int_t^T \Bigg[\frac{1}{2}\|\vartheta(s)\|^2 y^2 \mathfrak{X}_{yy}(s,y) + \big(\|\vartheta(s)\|^2 - r(s)\big)\, y\mathfrak{X}_y(s,y) - r(s)\mathfrak{X}(s,y)$$

$$- \vartheta^*(s)y\Psi_y^{\mathfrak{X}}(s,y) + \mu(s)I\big(s, y\mu(s)\big)\Bigg]ds - \int_t^T \big(\Psi^{\mathfrak{X}}(s,y)\big)^* \, dW_0(s)$$

for the pair $(\mathfrak{X}, \Psi^{\mathfrak{X}})$, namely, the semimartingale decomposition of the stochastic processes $\mathfrak{X}(\cdot, y)$ defined in (6.13) for each $y \in \mathbb{R}^+$.

The random field $\mathfrak{Y}(\cdot,\cdot)$ represents the random dynamic extension of the function $\mathcal{Y}(\cdot)$ introduced in section 5; in particular, $\mathcal{Y}(\cdot) = \mathfrak{Y}(0,\cdot)$.

Remark 6.8. Combining (2.7) with (8.8)–(8.11) of Proposition 8.3 in the appendix, with the identification of processes made in the proof of the preceding lemma, we obtain the dynamics

$$d\left[\beta(s)\,\mathfrak{X}\left(s, \frac{Y^{(0,y)}(s)}{\mu(s)}\right)\right] = -\beta(s)\Bigg\{\mu(s)I\big(s, Y^{(0,y)}(s)\big)\, ds$$

$$+ \left[\vartheta(s)\frac{Y^{(0,y)}(s)}{\mu(s)}\mathfrak{X}_y\left(s, \frac{Y^{(0,y)}(s)}{\mu(s)}\right) - \Psi^{\mathfrak{X}}\left(s, \frac{Y^{(0,y)}(s)}{\mu(s)}\right)\right]^* dW_0(s)\Bigg\},$$

and integrate to obtain

$$\beta(t)\,\mathfrak{X}\left(t, \frac{Y^{(0,y)}(t)}{\mu(t)}\right) + \int_0^t \beta(s)\mu(s)I\big(s, Y^{(0,y)}(s)\big)\,ds \tag{6.18}$$

$$= \mathfrak{X}(0,y) - \int_0^t \beta(s)\left[\vartheta(s)\frac{Y^{(0,y)}(s)}{\mu(s)}\,\mathfrak{X}_y\left(s, \frac{Y^{(0,y)}(s)}{\mu(s)}\right) - \Psi^{\mathfrak{X}}\left(s, \frac{Y^{(0,y)}(s)}{\mu(s)}\right)\right]^* dW_0(s)$$

for every $(t,y) \in [0,T] \times \mathbb{R}^+$, almost surely.

We are in a position now to obtain *stochastic feedback formulae* for the optimal investment and consumption processes. In view of (6.12), for each $t \in [0,T)$, the effective range for the running optimal wealth $X_0(t)$ and for the associated standard of living $z_0(t)$ will be

$$\mathcal{D}_t \triangleq \big\{(\mathfrak{x},\mathfrak{z}) \in \mathbb{R}^+ \times [0,\infty);\ \mathfrak{x} > w(t)\mathfrak{z}\big\}. \tag{6.19}$$

THEOREM 6.9. *Under Assumptions* 6.1 *and* 6.2*, the optimal consumption* $c_0(\cdot)$ *and the optimal portfolio strategy* $\pi_0(\cdot)$ *for the dynamic optimization problem* (4.8) *are given in the "stochastic adapted feedback form"* (6.1)*, where the random fields* C *and* Π *are given for* $t \in [0,T)$ *and* $(x,z) \in \mathcal{D}_t$ *as*

(6.20)
$$C(t,x,z) \triangleq z + I(t, \mu(t)\,\mathfrak{Y}(t, x - w(t)z)),$$

(6.21)
$$\Pi(t,x,z) \triangleq -\frac{1}{x}\big(\sigma^*(t)\big)^{-1}\left[\vartheta(t)\,\frac{\mathfrak{Y}\big(t, x-w(t)z\big)}{\mathfrak{Y}_x\big(t, x-w(t)z\big)} - \Psi^{\mathfrak{X}}\Big(t, \mathfrak{Y}\big(t, x-w(t)z\big)\Big)\right].$$

Proof. For any initial wealth x and standard of living z such that $(x,z) \in \mathcal{D}_0$ of (6.19), we may rewrite (6.12) as

$$Y^{(0,\mathcal{J})}(t)\Big|_{\mathcal{J}=\mathfrak{Y}(0,x-wz)} = \mu(t)\mathcal{J}(t) \qquad \text{with} \qquad \mathcal{J}(t) \triangleq \mathfrak{Y}\big(t, X_0(t) - w(t)z_0(t)\big).$$

From (5.5) and (6.8), it develops that the optimal consumption process of (5.10) is expressed by

$$c_0(t) = z_0(t) + I\big(t, \mu(t)\mathcal{J}(t)\big), \qquad 0 \le t < T,$$

and (6.20) is proved. Considering (6.18) for $y = \mathfrak{Y}(0, x - wz)$, in connection with (6.12), we obtain

$$\beta(t)\Big[X_0(t) - w(t)z_0(t)\Big] + \int_0^t \beta(s)\mu(s)\Big[c_0(s) - z_0(s)\Big]ds$$

$$= x - wz - \int_0^t \beta(s)\Big[\vartheta(s)\mathcal{J}(s)\,\mathfrak{X}_y\big(s, \mathcal{J}(s)\big) - \Psi^{\mathfrak{X}}\big(s, \mathcal{J}(s)\big)\Big]^* dW_0(s).$$

Now differentiate (6.16), to arrive at $\mathfrak{X}_y\big(t, \mathfrak{Y}(t, x - w(t)z)\big) = 1/\mathfrak{Y}_x(t, x - w(t)z)$ for every $(x,z) \in \mathcal{D}_t$; setting $\mathcal{J}_x(t) \triangleq \mathfrak{Y}_x\big(t, X_0(t) - w(t)z_0(t)\big)$ and using (6.6), the above equation becomes

$$\begin{aligned}\beta(t)X_0(t) &+ \int_0^t \beta(s)c_0(s)ds \\ &= x - \int_0^t \beta(s)\left[\vartheta(s)\frac{\mathcal{J}(s)}{\mathcal{J}_x(s)} - \Psi^{\mathfrak{X}}(s,\mathcal{J}(s))\right]^* dW_0(s) + \beta(t)w(t)z_0(t) \\ &\quad - wz - \int_0^t \beta(s)\delta(s)w(s)\Big[c_0(s) - z_0(s)\Big]ds + \int_0^t \beta(s)z_0(s)ds. \qquad (6.22)\end{aligned}$$

On the other hand, use (4.3) and (6.7) to compute

$$\begin{aligned}\beta(t)w(t)z_0(t) - wz &= \int_0^t d\Big(\beta(s)w(s)z_0(s)\Big) \\ &= \int_0^t \beta(s)\delta(s)w(s)\Big[c_0(s) - z_0(s)\Big]ds - \int_0^t \beta(s)z_0(s)ds, \qquad (6.23)\end{aligned}$$

and conclude that (6.22) reads

$$\beta(t)X_0(t) + \int_0^t \beta(s)c_0(s)ds = x - \int_0^t \beta(s)\left[\vartheta(s)\frac{\mathcal{J}(s)}{\mathcal{J}_x(s)} - \Psi^{\mathfrak{X}}(s,\mathcal{J}(s))\right]^* dW_0(s),$$

almost surely. A comparison of the latter with the integral expression (2.12) implies that (6.21) follows from $X_0(t)\pi_0^*(t)\sigma(t) = -\big[\vartheta(t)\frac{\mathcal{J}(t)}{\mathcal{J}_x(t)} - \Psi^{\mathfrak{X}}(t,\mathcal{J}(t))\big]^*$. □

Remark 6.10. Under the additional assumption of *deterministic* coefficients $r(\cdot) : [0,T] \to \mathbb{R}$, $\vartheta(\cdot) : [0,T] \to \mathbb{R}^d$, $\sigma(\cdot) : [0,T] \to L(\mathbb{R}^d;\mathbb{R}^d)$, $\alpha(\cdot) : [0,T] \to [0,\infty)$, and $\delta(\cdot) : [0,T] \to [0,\infty)$, the process $Y^{(t,y)}(\cdot)$ of (6.8) is Markovian. (Here $L(\mathbb{R}^d;\mathbb{R}^d)$ is the set of $(d \times d)$ matrices.) Thus, the random fields of (6.20) and (6.21), which represent the optimal policies in feedback form, reduce to the *deterministic functions*

$$C(t,x,z) = z + I\big(t, \mu(t)\mathcal{Y}(t, x - w(t)z)\big), \qquad (6.24)$$

$$\Pi(t,x,z) = -(\sigma^*(t))^{-1}\vartheta(t) \cdot \frac{\mathcal{Y}\big(t, x - w(t)z\big)}{x\mathcal{Y}_x\big(t, x - w(t)z\big)}, \qquad (6.25)$$

where $\mathcal{Y}(t,\cdot)$ is the inverse of the function

$$\mathcal{X}(t,y) \triangleq E^0\left[\int_t^T e^{-\int_t^s r(v)dv}\mu(s)I\big(s, yY^{(t,1)}(s)\big)ds\right], \qquad 0 < y < \infty;$$

cf. Lemma 6.6 and (6.13). It is then evident that the decision maker needs only to keep track of his current level of wealth $X_0(t)$ and standard of living $z_0(t)$, not of the entire history of the market up to time t; in other words, these quantities serve as *sufficient statistics* for the optimization problem (4.8).

7. The stochastic Hamilton–Jacobi–Bellman equation. We shall investigate now the analytical behavior of the value function for the optimization problem (4.8) as a solution of a nonlinear partial differential equation, widely referred to as the *stochastic Hamilton–Jacobi–Bellman equation.* In this vein, we find it useful to generalize the time horizon of our asset market $\mathcal{M}_0$ by taking initial date $t \in [0,T]$

rather than zero. Hence, for a fixed starting time $t \in [0,T]$ and any given capital wealth/initial standard of living pair $(x,z) \in \mathcal{D}_t$ (cf. (6.19)), the wealth process $X^{t,x,\pi,c}(\cdot)$, corresponding to a portfolio strategy $\pi(\cdot)$ and a consumption process $c(\cdot)$, satisfies the stochastic integral equation

$$X(s) = x + \int_t^s [r(v)X(v) - c(v)]dv + \int_t^s X(v)\pi^*(v)\sigma(v)dW_0(v) \tag{7.1}$$

for $t \le s \le T$, and the respective standard of living process $z(\cdot)$ is developed by

$$z(s) = ze^{-\int_t^s \alpha(\theta)d\theta} + \int_t^s \delta(v)e^{-\int_v^s \alpha(\theta)d\theta}c(v)dv, \quad t \le s \le T. \tag{7.2}$$

In this context, we shall call *admissible at the initial condition* (t,x), and denote their class by $\mathcal{A}(t,x)$, all portfolio/consumption pairs (π,c) such that $X^{t,x,\pi,c}(s) \ge 0$ for all $s \in [t,T]$, almost surely. Each of these pairs satisfies the budget constraint

$$E_t\left[\int_t^T H^t(s)c(s)ds\right] \le x. \tag{7.3}$$

Conversely, a variant of Lemma 2.1, subject to an initial date t that is not necessarily zero, shows that for every given consumption plan $c(\cdot)$ satisfying (7.3) we can fashion a portfolio strategy $\pi(\cdot)$ such that $(\pi,c) \in \mathcal{A}(t,x)$. Furthermore, we extend the optimization problem of Definition 4.2 by the random field

$$\mathrm{V}(t,x,z) \triangleq \operatorname*{ess\,sup}_{(\pi,c)\in\mathcal{A}'(t,x,z)} E_t\left[\int_t^T u\big(s, c(s) - z(s)\big)ds\right], \tag{7.4}$$

where

$$\mathcal{A}'(t,x,z) \triangleq \left\{(\pi,c) \in \mathcal{A}(t,x);\ E_t\left[\int_t^T u^-\big(s, c(s) - z(s)\big)ds\right] < \infty, \text{ almost surely}\right\}, \tag{7.5}$$

and $\mathrm{V}(0,\cdot,\cdot) = V(\cdot,\cdot)$. Invoking Assumptions 6.1 and 6.2, and imitating the methodology deployed in section 5, we obtain the almost sure representation

$$\mathrm{V}(t,x,z) = \mathfrak{G}\big(t, \mathfrak{Y}(t, x - w(t)z)\big), \quad (x,z) \in \mathcal{D}_t, \quad t \in [0,T), \tag{7.6}$$

by analogy with (5.16), where we have also introduced the real-valued random field

$$\mathfrak{G}(t,y) \triangleq E_t\left[\int_t^T u\big(s, I(s, yY^{(t,1)}(s))\big)ds\right], \qquad (t,y) \in [0,T] \times \mathbb{R}^+. \tag{7.7}$$

One observes that the random fields (7.4) and (7.7) constitute the dynamic, probabilistic analogues of those in (4.8) and (5.17), respectively, since $V(\cdot,\cdot) = \mathrm{V}(0,\cdot,\cdot)$ and $G(\cdot) = \mathfrak{G}(0,\cdot)$; this complies with the temporal and stochastic evolution of the function $\mathcal{X}(\cdot)$ described in the previous section. Clearly

$$\mathrm{V}(T,x,z) = 0 \ \ \forall\ (x,z) \in \mathcal{D}; \tag{7.8}$$

in fact, $\mathrm{V}(t,x,z) < \infty$ for every $t \in [0,T)$, $(x,z) \in \mathcal{D}_t$, and with $\partial\mathcal{D}_t = \big\{(\mathfrak{x},\mathfrak{z}) \in [0,\infty)^2;\ \mathfrak{x} = w(t)\mathfrak{z}\big\}$ the boundary of $\mathcal{D}_t$ (cf. (5.25)) we have

$$\lim_{\substack{(x,z)\to(\chi,\zeta)\\(x,z)\in\mathcal{D}_t}} \mathrm{V}(t,x,z) = \int_t^T u(s,0^+)ds \quad \forall\ (\chi,\zeta) \in \partial\mathcal{D}_t. \tag{7.9}$$

We derive now a semimartingale decomposition for the random field $\mathfrak{G}$ of (7.7).

LEMMA 7.1. *Under Assumptions* 6.1 *and* 6.2, *there exists a random field* $\Phi^{\mathfrak{G}} : [0,T]\times\mathbb{R}^+\times\Omega \to \mathbb{R}^d$ *such that the pair of random fields* $(\mathfrak{G},\Phi^{\mathfrak{G}})$, *where* $\mathfrak{G}$ *is given by* (7.7), *is of class* $C_{\mathbb{F}}\big([0,T];\mathbb{L}^2(\Omega;C^3(\mathbb{R}^+))\big)\times\mathbb{L}^2_{\mathbb{F}}\big(0,T;\mathbb{L}^2(\Omega;C^2(\mathbb{R}^+;\mathbb{R}^d))\big)$ *and provides the unique solution of the Cauchy problem*

$$\begin{aligned}-d\mathfrak{G}(t,y) = \Bigg[&\frac{1}{2}\|\vartheta(t)\|^2y^2\mathfrak{G}_{yy}(t,y) - r(t)y\mathfrak{G}_y(t,y)\\ &- \vartheta^*(t)y\Phi^{\mathfrak{G}}_y(t,y) + u\big(t,I(t,y\mu(t))\big)\Bigg]dt - \big(\Phi^{\mathfrak{G}}(t,y)\big)^*\,dW(t)\end{aligned} \tag{7.10}$$

on $[0,T)\times\mathbb{R}^+$, *and the terminal condition*

$$\mathfrak{G}(T,y) = 0 \quad \textit{on } \mathbb{R}^+, \tag{7.11}$$

almost surely. Moreover, for every $(t,y) \in [0,T)\times\mathbb{R}^+$ *we have almost surely*

$$\mathfrak{G}(t,y) - \mathfrak{G}(t,z) = y\mathfrak{X}(t,y) - z\mathfrak{X}(t,z) - \int_z^y \mathfrak{X}(t,\xi)d\xi, \quad 0 < z < y < \infty, \tag{7.12}$$

$$\mathfrak{G}_y(t,y) = y\mathfrak{X}_y(t,y), \qquad \mathfrak{G}_{yy}(t,y) = \mathfrak{X}_y(t,y) + y\mathfrak{X}_{yy}(t,y). \tag{7.13}$$

Once again (cf. Remark 6.7), the additional smoothness of $(\mathfrak{G},\Phi^{\mathfrak{G}})$ will be essential for the formal derivation of explicit calculations, and the semimartingale decomposition of the process $\mathfrak{G}(\cdot,y)$, $y \in \mathbb{R}^+$ is given as

$$\begin{aligned}\mathfrak{G}(t,y) = \int_t^T \Bigg[&\frac{1}{2}\|\vartheta(s)\|^2y^2\mathfrak{G}_{yy}(s,y) - r(s)y\mathfrak{G}_y(s,y)\\ &- \vartheta^*(s)y\Phi^{\mathfrak{G}}_y(s,y) + u\big(s,I(s,y\mu(s))\big)\Bigg]ds - \int_t^T \big(\Phi^{\mathfrak{G}}(s,y)\big)^*\,dW(s).\end{aligned}$$

Proof of Lemma 7.1. Use (6.9), (6.11), and (7.7), and apply Proposition 8.3 in the appendix for $m(\cdot) = (\mu'(\cdot)/\mu(\cdot)) - r(\cdot)$, $\nu(\cdot) = -\vartheta(\cdot)$, $N(\cdot,\cdot) = u\big(\cdot,I(\cdot,\cdot)\big)$, $\ell(\cdot) = \mu(\cdot)$, $q(\cdot) = 1$, and $\rho(\cdot) = 0$, to check that the pair of random fields $(\mathfrak{G},\Phi^{\mathfrak{G}})$ has the asserted order of regularity and is the unique solution of the Cauchy problem (7.10), (7.11), almost surely. Repeat the computations in (5.33) concerning conditional expectations, subject to an initial time $t \neq 0$, to obtain (7.12); differentiation then yields (7.13). □

We derive now the semimartingale decomposition of the random field $\mathfrak{Y}$ in (6.16).

LEMMA 7.2. *Under the hypotheses of Lemma* 6.6, *there exists a pair of random fields* $(\Theta,\Sigma) \in \mathbb{L}_{\mathbb{F}}\big(0,T';C^1(\mathbb{R}^+)\big)\times\mathbb{L}^2_{\mathbb{F}}\big(0,T';C^2(\mathbb{R}^+;\mathbb{R}^d)\big)$ *for each* $0 < T' < T$, *such that*

$$d\,\mathfrak{Y}(t,x) = -\Theta(t,x)dt + \Sigma^*(t,x)dW_0(t) \tag{7.14}$$

holds almost surely for every $(t,x) \in [0,T) \times \mathbb{R}^+$. *In particular, these random fields are uniquely determined by the following relationships:*

$$\begin{aligned}
&\frac{1}{2}\Big[\|\Sigma(t,x)\|^2 - \|\vartheta(t)\|^2\mathfrak{Y}^2(t,x)\Big]\mathfrak{X}_{yy}\big(t,\mathfrak{Y}(t,x)\big) - \mu(t)\, I\big(t,\mu(t)\,\mathfrak{Y}(t,x)\big)\\
&\quad + \Big[\big(r(t) - \|\vartheta(t)\|^2\big)\,\mathfrak{Y}(t,x) + \vartheta^*(t)\Sigma(t,x) - \Theta(t,x)\Big]\mathfrak{X}_y\big(t,\mathfrak{Y}(t,x)\big)\\
&\quad + r(t)x + \Big[\Sigma(t,x) + \vartheta(t)\mathfrak{Y}(t,x)\Big]^*\Psi^{\mathfrak{X}}_y\big(t,\mathfrak{Y}(t,x)\big) + \vartheta^*(t)\Psi^{\mathfrak{X}}\big(t,\mathfrak{Y}(t,x)\big) = 0
\end{aligned} \tag{7.15}$$

and

$$\mathfrak{X}_y\big(t,\mathfrak{Y}(t,x)\big)\,\Sigma(t,x) + \Psi^{\mathfrak{X}}\big(t,\mathfrak{Y}(t,x)\big) = 0. \tag{7.16}$$

Proof. Let $(t,x) \in [0,T) \times \mathbb{R}^+$. Invoking (6.14) for $\mathfrak{X}$ and postulating the representation (7.14) for $\mathfrak{Y}$, we differentiate the identity (6.16) with respect to the temporal variable, using Proposition 8.1 in the appendix and (2.9), and then integrate over $[0,t]$ to compute

$$\begin{aligned}
&\int_0^t \bigg\{\frac{1}{2}\Big[\|\Sigma(s,x)\|^2 - \|\vartheta(s)\|^2\mathfrak{Y}^2(s,x)\Big]\mathfrak{X}_{yy}(s,\mathfrak{Y}(s,x)) - \mu(s)I\big(s,\mu(s)\mathfrak{Y}(s,x)\big)\\
&\qquad + \Big[\big(r(s) - \|\vartheta(s)\|^2\big)\,\mathfrak{Y}(s,x) + \vartheta^*(s)\Sigma(s,x) - \Theta(s,x)\Big]\mathfrak{X}_y(s,\mathfrak{Y}(s,x))\\
&\qquad + r(s)x + \Big[\Sigma(s,x) + \vartheta(s)\mathfrak{Y}(s,x)\Big]^*\Psi^{\mathfrak{X}}_y\big(s,\mathfrak{Y}(s,x)\big) + \vartheta^*(s)\Psi^{\mathfrak{X}}\big(s,\mathfrak{Y}(s,x)\big)\bigg\}ds\\
&+ \int_0^t \Big\{\mathfrak{X}_y\big(s,\mathfrak{Y}(s,x)\big)\Sigma(s,x) + \Psi^{\mathfrak{X}}\big(s,\mathfrak{Y}(s,x)\big)\Big\}^* dW(s) = 0,
\end{aligned}$$

almost surely. Thus, the uniqueness for the decomposition of a continuous semimartingale (see, e.g., Karatzas and Shreve (1991), page 149) implies that both integrals of the above equation vanish. Differentiation of the Lebesgue integral implies (7.15), while the quadratic variation of the stochastic integral vanishes as well, leading to (7.16). The derived equations define uniquely the random fields Θ and Σ, assigning to them the claimed order of adaptivity, integrability, and smoothness; these are then seen to satisfy the representation (7.14). □

LEMMA 7.3. *Under Assumptions* 6.1 *and* 6.2, *the random fields* $\Psi^{\mathfrak{X}}$ *of* (6.14) *and* $\Phi^{\mathfrak{G}}$ *of* (7.10) *satisfy almost surely the relationship*

$$\Phi^{\mathfrak{G}}_y(t,y) - y\Psi^{\mathfrak{X}}_y(t,y) = 0 \quad \forall\ (t,y) \in [0,T) \times \mathbb{R}^+. \tag{7.17}$$

Proof. Taking time differentials in (7.12), we get almost surely

$$d\mathfrak{G}(t,y) - d\mathfrak{G}(t,z) = y\,d\mathfrak{X}(t,y) - z\,d\mathfrak{X}(t,z) - \int_z^y d\mathfrak{X}(t,\lambda)\,d\lambda, \quad 0 \le t < T.$$

Now make the substitutions (6.14), (7.10) in the above formula, and equate the respective martingale parts (see, e.g., Karatzas and Shreve (1991), Problem 3.3.2) to obtain

$$\Phi^{\mathfrak{G}}(t,y) - \Phi^{\mathfrak{G}}(t,z) = y\Psi^{\mathfrak{X}}(t,y) - z\,\Psi^{\mathfrak{X}}(t,z) - \int_z^y \Psi^{\mathfrak{X}}(t,\lambda)\,d\lambda. \tag{7.18}$$

Of course, (7.18) is valid only if the interchange of Lebesgue and Itô integrals

$$\int_z^y \int_0^t \Psi^{\mathfrak{X}}(s,\lambda)dW(s)\ d\lambda = \int_0^t \int_z^y \Psi^{\mathfrak{X}}(s,\lambda)d\lambda\ dW(s)$$

holds almost surely for each $t \in [0,T)$. But this is true, due to the observation that $L(t,\cdot) = \int_z^{\cdot} \Psi^{\mathfrak{X}}(t,\lambda)d\lambda$ is a C^2 random field on $[z,\infty)$ and Exercise 3.1.5 in Kunita (1990). Differentiating (7.18) we obtain (7.17). □

We are ready now to state the main result of this section.

Theorem 7.4 (stochastic Hamilton–Jacobi–Bellman equation). *Under Assumptions* 6.1 *and* 6.2*, the pair of random fields* (V,Ξ) *given by* (7.6)–(7.8) *and*

$$\Xi(t,x,z) \triangleq \Phi^{\mathfrak{G}}\big(t,\mathfrak{Y}(t,x-w(t)z)\big) - \mathfrak{Y}\big(t,x-w(t)z\big)\,\Psi^{\mathfrak{X}}\big(t,\mathfrak{Y}(t,x-w(t)z)\big) \tag{7.19}$$

is of class

$$C_{\mathbb{F}}\big(\big\{t\in[0,T];\ V(t,\cdot,\cdot)\in C^{3,3}(\mathcal{D}_t)\big\}\big)\times \mathbb{L}^2_{\mathbb{F}}\big(\big\{t\in[0,T);\ \Xi(t,\cdot,\cdot)\in C^{2,2}(\mathcal{D}_t;\mathbb{R}^d)\big\}\big)$$

and satisfies on $\big\{(t,x,z);\ t\in[0,T),\ (x,z)\in\mathcal{D}_t\big\}$ *the* stochastic Hamilton–Jacobi–Bellman *partial differential equation of dynamic programming*

$$\begin{aligned} -d\mathrm{V}(t,x,z) = \operatorname*{ess\,sup}_{\substack{0\le c<\infty\\ \pi\in\mathbb{R}^d}}\Bigg\{&\frac{1}{2}\|\sigma^*(t)\pi\|^2x^2\mathrm{V}_{xx}(t,x,z)\\ &+\Big[r(t)x-c+\pi^*\sigma(t)\vartheta(t)x\Big]\mathrm{V}_x(t,x,z)\\ &+\Big[\delta(t)c-\alpha(t)z\Big]\mathrm{V}_z(t,x,z)+\pi^*\sigma(t)x\Xi_x(t,x,z)\\ &+u(t,c-z)\Bigg\}dt-\Xi(t,x,z)dW(t)\end{aligned} \tag{7.20}$$

as well as the boundary conditions (7.8) *and* (7.9)*.*

Furthermore, the random fields $\Pi(t,x,z), C(t,x,z)$ *of* (6.20), (6.21) *provide the optimal portfolio/consumption rules for the maximization in* (7.20)*.*

Proof. Differentiation of (6.16), (7.6), and (7.19), in combination with (7.13) and (7.17), leads almost surely to

$$\mathfrak{X}_y\big(t,\mathfrak{Y}(t,x-w(t)z)\big)\,\mathfrak{Y}_x\big(t,x-w(t)z\big)=1,$$

$$\mathrm{V}_x(t,x,z)\ =\ \mathfrak{Y}\big(t,x-w(t)z\big),\quad \mathrm{V}_z(t,x,z)=-w(t)\,\mathfrak{Y}\big(t,x-w(t)z\big),$$

$$\mathrm{V}_{xx}(t,x,z)\ =\ \mathfrak{Y}_x(t,x-w(t)z),\Xi_x(t,x,z)=-\mathfrak{Y}_x\big(t,x-w(t)z\big)\,\Psi^{\mathfrak{X}}\big(t,\mathfrak{Y}(t,x-w(t)z)\big)$$

for $(x,z)\in\mathcal{D}_t$, $0\le t<T$. Using these formulae and (6.6), the right-hand side of (7.20) becomes

$$
\begin{aligned}
&\Big[r(t)x\,\mathfrak{Y}(t,x-w(t)z)+\alpha(t)w(t)z\,\mathfrak{Y}(t,x-w(t)z)\\
&\quad+\operatorname*{ess\,sup}_{0\le c<\infty}\big\{u(t,c-z)-c\,\mu(t)\,\mathfrak{Y}(t,x-w(t)z)\big\}\\
&\quad+\operatorname*{ess\,sup}_{\pi\in\mathbb{R}^d}\Big\{\frac{1}{2}\|\sigma^*(t)\pi\|^2x^2\,\mathfrak{Y}_x(t,x-w(t)z)+\pi^*\sigma(t)x\Big[\vartheta(t)\mathfrak{Y}(t,x-w(t)z)\\
&\qquad\qquad\qquad-\mathfrak{Y}_x(t,x-w(t)z)\Psi^{\mathfrak{X}}\big(t,\mathfrak{Y}(t,x-w(t)z)\big)\Big]\Big\}\Big]\,dt\\
&-\Big[\Phi^{\mathfrak{G}}\big(t,\mathfrak{Y}(t,x-w(t)z)\big)-\mathfrak{Y}\big(t,x-w(t)z\big)\,\Psi^{\mathfrak{X}}\big(t,\mathfrak{Y}(t,x-w(t)z)\big)\Big]dW(t).
\end{aligned}
$$

The strict concavity and smoothness of both expressions to be maximized allow us to differentiate and solve the resulting first-order equations, in order to obtain the optimal values of c and π. These values turn out to coincide with (6.20) and (6.21), respectively. Substituting them now into the latter expression, we are led to

$$
\begin{aligned}
&\Big[r(t)x\,\mathfrak{Y}\big(t,x-w(t)z\big)+\alpha(t)w(t)z\,\mathfrak{Y}\big(t,x-w(t)z\big)\\
&\quad+u\Big(t,I\big(t,\mu(t)\mathfrak{Y}(t,x-w(t)z)\big)\Big)\\
&\quad-\mu(t)\,\mathfrak{Y}\big(t,x-w(t)z\big)\Big[z+I\big(t,\mu(t)\mathfrak{Y}(t,x-w(t)z)\big)\Big]\\
&\quad-\frac{1}{2\mathfrak{Y}_x(t,x-w(t)z)}\Big\|\vartheta(t)\,\mathfrak{Y}(t,x-w(t)z)\\
&\qquad\qquad-\mathfrak{Y}_x(t,x-w(t)z)\Psi^{\mathfrak{X}}\big(t,\mathfrak{Y}(t,x-w(t)z)\big)\Big\|^2\Big]\,dt\\
&-\Big[\Phi^{\mathfrak{G}}\big(t,\mathfrak{Y}(t,x-w(t)z)\big)-\mathfrak{Y}\big(t,x-w(t)z\big)\,\Psi^{\mathfrak{X}}\big(t,\mathfrak{Y}(t,x-w(t)z)\big)\Big]^*dW(t).
\end{aligned}
\tag{7.21}
$$

On the other hand, couple (7.14) with (2.9), and apply the generalized Itô–Kunita–Wentzell formula of Proposition 8.1, to derive the representation

$$
\begin{aligned}
&d\,\mathfrak{Y}(t,x-w(t)z)\\
&\quad=\Big[\vartheta^*(t)\Sigma(t,x-w(t)z)-\Theta(t,x-w(t)z)-w'(t)z\,\mathfrak{Y}_x\big(t,x-w(t)z\big)\Big]dt\\
&\qquad+\Sigma^*(t,x-w(t)z)\,dW(t).
\end{aligned}
$$

Using this, we apply the Itô–Kunita–Wentzell formula a second time, now involving (7.6) and (7.10), and obtain the left-hand side of (7.20) as

$$\begin{aligned}
&\bigg[-\frac{1}{2}\Big[\|\Sigma(t,x-w(t)z)\|^2-\|\vartheta(t)\|^2\mathfrak{Y}^2(t,x-w(t)z)\Big]\,\mathfrak{G}_{yy}\big(t,\mathfrak{Y}(t,x-w(t)z)\big)\\
&\quad-\Big[r(t)\mathfrak{Y}(t,x-w(t)z)+\vartheta^*(t)\Sigma(t,x-w(t)z)-\Theta(t,x-w(t)z)\\
&\qquad\qquad-w'(t)z\,\mathfrak{Y}_x\big(t,x-w(t)z\big)\Big]\mathfrak{G}_y\big(t,\mathfrak{Y}(t,x-w(t)z)\big)\\
&\quad-\Big[\Sigma(t,x-w(t)z)+\vartheta(t)\mathfrak{Y}(t,x-w(t)z)\Big]^*\Phi^{\mathfrak{G}}_y\big(t,\mathfrak{Y}(t,x-w(t)z)\big)\\
&\quad+u\Big(t,I\big(t,\mu(t)\mathfrak{Y}(t,x-w(t)z)\big)\Big)\bigg]dt\\
&-\Big[\Phi^{\mathfrak{G}}\big(t,\mathfrak{Y}(t,x-w(t)z)\big)+\mathfrak{G}_y\big(t,\mathfrak{Y}(t,x-w(t)z)\big)\Sigma(t,x-w(t)z)\Big]^*dW(t),
\end{aligned}$$

which via (7.13) becomes

$$\begin{aligned}
&\bigg[-\frac{1}{2}\|\Sigma(t,x-w(t)z)\|^2\mathfrak{X}_y\big(t,\mathfrak{Y}(t,x-w(t)z)\big)\\
&\quad-\mathfrak{Y}(t,x-w(t)z)\bigg\{\frac{1}{2}\Big[\|\Sigma(t,x-w(t)z)\|^2\\
&\qquad\qquad\qquad-\|\vartheta(t)\|^2\mathfrak{Y}^2(t,x-w(t)z)\Big]\mathfrak{X}_{yy}\big(t,\mathfrak{Y}(t,x-w(t)z)\big)\\
&\qquad-\Big[r(t)\mathfrak{Y}(t,x-w(t)z)-\frac{1}{2}\|\vartheta(t)\|^2\mathfrak{Y}(t,x-w(t)z)+\vartheta^*(t)\Sigma(t,x-w(t)z)\\
&\qquad\qquad-\Theta(t,x-w(t)z)-w'(t)z\,\mathfrak{Y}_x\big(t,x-w(t)z\big)\Big]\mathfrak{X}_y\big(t,\mathfrak{Y}(t,x-w(t)z)\big)\bigg\}\\
&\quad-\Big[\Sigma(t,x-w(t)z)+\vartheta(t)\mathfrak{Y}(t,x-w(t)z)\Big]^*\Phi^{\mathfrak{G}}_y\big(t,\mathfrak{Y}(t,x-w(t)z)\big)\\
&\quad+u\Big(t,I\big(t,\mu(t)\mathfrak{Y}(t,x-w(t)z)\big)\Big)\bigg]dt\\
&-\Big[\Phi^{\mathfrak{G}}\big(t,\mathfrak{Y}(t,x-w(t)z)\big)\\
&\qquad+\mathfrak{Y}(t,x-w(t)z)\mathfrak{X}_y\big(t,\mathfrak{Y}(t,x-w(t)z)\big)\Sigma(t,x-w(t)z)\Big]^*dW(t).
\end{aligned}$$

Finally, Lemmata 7.2 and 7.3 transform the latter to

$$\begin{aligned}
&\bigg[\Big[r(t)\big(x-w(t)z\big)+w'(t)z\Big]\mathfrak{Y}(t,x-w(t)z)+u\Big(t,I\big(t,\mu(t)\mathfrak{Y}(t,x-w(t)z)\big)\Big)\\
&\quad+\Big[\vartheta^*(t)\Psi^{\mathfrak{X}}\big(t,\mathfrak{Y}(t,x-w(t)z)\big)-\mu(t)I\big(t,\mu(t)\mathfrak{Y}(t,x-w(t)z)\big)\Big]\mathfrak{Y}(t,x-w(t)z)
\end{aligned}$$

$$- \frac{1}{2}\|\vartheta(t)\|^2 \frac{\mathfrak{Y}^2(t, x - w(t)z)}{\mathfrak{Y}_x(t, x - w(t)z)} - \frac{1}{2}\|\Psi^{\mathfrak{X}}\big(t, \mathfrak{Y}(t, x - w(t)z)\big)\|^2 \mathfrak{Y}_x(t, x - w(t)z)\Bigg] dt$$

$$- \Big[\Phi^{\mathfrak{G}}\big(t, \mathfrak{Y}(t, x - w(t)z)\big) - \mathfrak{Y}(t, x - w(t)z)\Psi^{\mathfrak{X}}\big(t, \mathfrak{Y}(t, x - w(t)z)\big)\Big]^* dW(t).$$

Expanding the norm in (7.21) and recalling (6.7), we conclude that both sides of (7.20) coincide almost surely. □

Remark 7.5. Carrying out the maximization according to the proof of Theorem 7.4, (7.20) takes the more conventional form

$$\begin{aligned} &d\,\mathrm{V}(t,x,z) + \mathbf{H}\Big(\mathrm{V}_{xx}(t,x,z), \mathrm{V}_x(t,x,z), \mathrm{V}_z(t,x,z), \Xi_x(t,x,z), t, x, z\Big) dt \\ &\qquad\qquad - \Xi(t,x,z)\, dW(t) = 0, \end{aligned} \tag{7.22}$$

where we denote by $\mathbf{H}$ the Hamiltonian

$$\begin{aligned} \mathbf{H}(\mathrm{A}, p, q, \mathrm{B}, t, x, z) \triangleq &- \frac{1}{2\mathrm{A}}\|\vartheta(t)p + \mathrm{B}\|^2 + \Big[r(t)x - z - I(t, p - \delta(t)q)\Big]p \\ &+ \Big[(\delta(t) - \alpha(t))z + \delta(t)I(t, p - \delta(t)q)\Big]q + u\big(t, I(t, p - \delta(t)q)\big) \end{aligned}$$

for $\mathrm{A} < 0$, $p > 0$, $q < 0$, and $\mathrm{B} \in \mathbb{R}$. Notice that we have obtained a closed-form solution of the *strongly nonlinear* stochastic Hamilton–Jacobi–Bellman equation (7.22), by solving instead the two *linear* equations (6.14), (7.10) subject to the appropriate initial and regularity conditions and then composing as in (7.6).

Remark 7.6. Theorem 7.4 provides a rare illustration of the Peng (1992) approach to stochastic Hamilton–Jacobi–Bellman equations. More precisely, it formulates the nonlinear SPDE satisfied by the value random field of the stochastic optimal control problem (7.4). To our knowledge, this is the first concrete illustration of BSPDEs in a stochastic control context beyond the classical linear/quadratic regulator worked out in Peng (1992).

As a consequence, (7.20) provides a *necessary condition* that must be satisfied by the value random field V of (7.4). Due to the absence of an appropriate growth condition for V as each component of $(x, z) \in \mathcal{D}_t$ increases to infinity, (7.20) fails to also be sufficient; in other words, we cannot claim directly that V is the unique solution of (7.20) with boundary conditions (7.8), (7.9). We decide though to treat this matter by establishing a *necessary and sufficient condition* for the *convex dual* of V, defined as

$$\widetilde{\mathrm{V}}(t, y) \triangleq \operatorname*{ess\,sup}_{(x,z)\in\mathcal{D}_t} \Big\{\mathrm{V}(t,x,z) - \big(x - w(t)z\big)y\Big\}, \qquad y \in \mathbb{R}, \tag{7.23}$$

by analogy with (5.24). Doing so, we avoid investigating the solvability of the *nonlinear* SPDE (7.20), since it turns out that $\widetilde{V}$ is equivalently characterized as the unique solution of a *linear* parabolic BSPDE (cf. (7.29), (7.30)) and V can be easily recovered by inverting the above Legendre–Fenchel transformation to have almost surely

$$\mathrm{V}(t,x,z) = \operatorname*{ess\,inf}_{y\in\mathbb{R}} \Big\{\widetilde{\mathrm{V}}(t,y) + \big(x - w(t)z\big)y\Big\}, \qquad (x,z) \in \mathcal{D}_t.$$

We formalize these considerations as follows.

THEOREM 7.7 (convex dual of $\mathrm{V}(t,\cdot)$). *Under Assumptions* 6.1 *and* 6.2*, for each given* $t \in [0,T)$ *the function* $\mathrm{V}(t,\cdot,\cdot)$ *is a generalized utility function, as defined in Theorem* 5.8*, almost surely; also,*

$$\mathrm{V}_x(t,x,z) = \mathfrak{Y}\big(t, x - w(t)z\big) \quad \forall\ (x,z) \in \mathcal{D}_t, \tag{7.24}$$

$$\mathrm{V}_z(t,x,z) = -w(t)\,\mathfrak{Y}(t, x - w(t)z) \quad \forall\ (x,z) \in \mathcal{D}_t. \tag{7.25}$$

Furthermore, for $(t,y) \in [0,T] \times \mathbb{R}^+$*, we have*

$$\widetilde{\mathrm{V}}(t,y) = \mathfrak{G}(t,y) - y\mathfrak{X}(t,y) = E_t\left[\int_t^T \widetilde{u}\big(s, yY^{(t,1)}(s)\big)\,ds\right], \tag{7.26}$$

$$\widetilde{\mathrm{V}}_y(t,y) = -\mathfrak{X}(t,y), \tag{7.27}$$

almost surely. Finally, the pair of random fields $(\widetilde{\mathrm{V}}, \Lambda)$*, where*

$$\Lambda(t,y) \triangleq \Phi^{\mathfrak{G}}(t,y) - y\Psi^{\mathfrak{X}}(t,y), \quad (t,y) \in [0,T] \times \mathbb{R}^+, \tag{7.28}$$

belongs to $C_{\mathbb{F}}\big([0,T];\mathbb{L}^2(\Omega;C^3(\mathbb{R}^+))\big) \times \mathbb{L}^2_{\mathbb{F}}\big(0,T;\mathbb{L}^2(\Omega;C^2(\mathbb{R}^+;\mathbb{R}^d))\big)$ *and is the unique solution of the following Cauchy problem for the* linear *BSPDE:*

$$\begin{aligned} -d\widetilde{\mathrm{V}}(t,y) = \Bigg[&\frac{1}{2}\|\vartheta(t)\|^2 y^2 \widetilde{\mathrm{V}}_{yy}(t,y) - r(t)y\widetilde{\mathrm{V}}_y(t,y) - \vartheta^*(t)y\Lambda_y(t,y) \\ &+ \widetilde{u}\big(t, y\mu(t)\big)\Bigg]dt - \Lambda^*(t,y)dW(t) \qquad \textit{on } [0,T) \times \mathbb{R}^+, \end{aligned} \tag{7.29}$$

$$\widetilde{\mathrm{V}}(T,y) = 0 \quad \textit{on } \mathbb{R}^+. \tag{7.30}$$

Merging now (7.19) and (7.28), we notice that the random fields Ξ and Λ of the martingale parts of V and $\widetilde{\mathrm{V}}$, respectively, are related via the almost sure expression

$$\Xi(t,x,z) = \Lambda\big(t, \mathfrak{Y}(t, x - w(t)z)\big), \quad t \in [0,T), \ \ (x,z) \in \mathcal{D}_t. \tag{7.31}$$

Proof of Theorem 7.7. Setting claim (5.25) aside, the first two parts of this result represent the dynamic, stochastic counterpart of Theorem 5.8. Thus, all the respective assertions, including (7.24)–(7.27), can be proved through a similar methodology, keeping in mind the new feature of conditional expectation. From Lemmata 6.6 and 7.1, (7.26), and (7.28), it is easy to verify the stated regularity for the pair $(\widetilde{\mathrm{V}}, \Lambda)$, while (7.29) and (7.30) are direct implications of (7.26), (7.10), (3.4), (6.14), and (7.11) with (6.15). □

Remark 7.8. In a Markovian framework with nonrandom model coefficients (cf. Remark 6.10), the unique solutions (6.13), (7.7), (7.4), and (7.23) of the SPDEs of Lemmata 6.6 and 7.1 and Theorems 7.4 and 7.7, respectively, are deterministic functions. In particular, the stochastic integrals in these equations vanish, reducing them to deterministic ones.

The example that follows illustrates the use of Theorem 7.7 as an alternative method for characterizing, even computing, the value random field and the stochastic feedback formulae of the optimal portfolio/consumption pair.

Example 7.9 (*logarithmic utility*). Take $u(t,x) = \log x$ for all $(t,x) \in [0,T] \times \mathbb{R}^+$; thus, $I(t,y) = 1/y$, $\tilde{u}(t,y) = -\log y - 1$ for $(t,y) \in [0,T] \times \mathbb{R}^+$.

Case 1: *Deterministic coefficients.* The Cauchy problem (7.29) now takes the form

$$\widetilde{V}_t(t,y) + \frac{1}{2}\|\vartheta(t)\|^2 y^2 \widetilde{V}_{yy}(t,y) - r(t) y \widetilde{V}_y(t,y) = -\tilde{u}(t, y\mu(t)) \quad \text{on } [0,T) \times \mathbb{R}^+. \tag{7.32}$$

Motivated by the nonhomogeneous term of (7.32), we seek appropriate functions $\nu, m : [0,T] \to \mathbb{R}$ such that

$$\tilde{v}(t,y) \triangleq -\nu(t)\log(y\mu(t)) - m(t) \tag{7.33}$$

satisfies (7.32), (7.30). Indeed, this is the case if and only if

$$\nu(t) = T - t, \qquad m(t) = \int_t^T \left[1 - (T-s)\left(\frac{1}{2}\|\vartheta(s)\|^2 + r(s) - \frac{\mu'(s)}{\mu(s)}\right)\right] ds \tag{7.34}$$

for $0 \le t \le T$, and then $\tilde{v} \in C([0,T] \times \mathbb{R}^+) \cap C^{1,3}([0,T) \times \mathbb{R}^+)$. From Theorem 7.7, $\tilde{v}$ is the unique solution of the Cauchy problem (7.32), (7.30); thus $\widetilde{V} \equiv \tilde{v}$,

$$\mathcal{X}(t,y) = \frac{\nu(t)}{y}, \quad G(t,y) = \nu(t)\Big[1 - \log(y\mu(t))\Big] - m(t), \quad (t,y) \in [0,T] \times \mathbb{R}^+.$$

Therefore,

$$\mathcal{Y}(t,x) = \frac{\nu(t)}{x},\ x \in \mathbb{R}^+, \quad V(t,x,z) = \nu(t)\log\left(\frac{x - w(t)z}{\nu(t)\mu(t)}\right) + \nu(t) - m(t),\ (x,z) \in \mathcal{D}_t,$$

and the feedback formulae (6.24), (6.25) for the optimal consumption and portfolio are given, for every $0 \le t < T$, by

$$C(t,x,z) = z + \frac{x - w(t)z}{\nu(t)\mu(t)} \quad \text{and} \quad \Pi(t,x,z) = (\sigma^*(t))^{-1}\vartheta(t)\frac{x - w(t)z}{x}, \quad (x,z) \in \mathcal{D}_t.$$

Case 2: *Random coefficients.* Our goal is to find an $\mathbb{F}$-*adapted* pair of random fields that satisfies (7.29), (7.30). By analogy with (7.33)–(7.34), we introduce in this case the $\mathbb{F}$-adapted random field

$$\tilde{\mathfrak{v}}(t,y) \triangleq -\nu(t)\log(y\mu(t)) - \mathfrak{m}(t)$$

for $(t,y) \in [0,T] \times \mathbb{R}^+$, with $\nu(t) = T - t$ and

$$\mathbf{m}(t) = E_t\left[\int_t^T \left\{1 - (T-s)\left(\frac{1}{2}\|\vartheta(s)\|^2 + r(s) - \frac{\mu'(s)}{\mu(s)}\right)\right\} ds\right].$$

Moreover, the completeness of the market stipulates the existence of an $\mathbb{R}^d$-valued, $\mathbb{F}$-progressively measurable, square-integrable process $\ell(\cdot)$, such that the Brownian martingale

$$\mathfrak{M}(t) = E_t\left[\int_0^T \left\{1 - (T-s)\left(\frac{1}{2}\|\vartheta(s)\|^2 + r(s) - \frac{\mu'(s)}{\mu(s)}\right)\right\} ds\right]$$

has the representation $\mathfrak{M}(t) = \mathfrak{M}(0) + \int_0^t \ell^*(s)\, dW(s)$, $0 \le t \le T$.

It is verified directly that the pair $(\tilde{\mathfrak{v}}, \ell)$, where $\tilde{\mathfrak{v}} \in C_{\mathbb{F}}\big([0,T]; \mathbb{L}^2(\Omega; C^3(\mathbb{R}^+))\big)$, satisfies (7.29), (7.30). Therefore, Theorem 7.7 implies that $(\widetilde{\mathfrak{v}}, \ell)$ agrees with $(\widetilde{\mathrm{V}}, \Lambda)$, and

$$\mathfrak{X}(t,y) = \frac{\nu(t)}{y}, \quad \mathfrak{G}(t,y) = \nu(t)\big[1 - \log(y\mu(t))\big] - \mathfrak{m}(t), \quad (t,y) \in [0,T] \times \mathbb{R}^+.$$

Consequently, for $0 \le t < T$, it transpires that $\mathfrak{Y}(t,x) = \nu(t)/x$, $x \in \mathbb{R}^+$, and

$$\mathrm{V}(t,x,z) = \nu(t) \log\left(\frac{x - w(t)z}{\nu(t)\mu(t)}\right) + \nu(t) - \mathfrak{m}(t), \quad (x,z) \in \mathcal{D}_t.$$

For this special choice of utility preference, $\mathfrak{X}$ (and so $\mathfrak{Y}$) is deterministic, and the feedback formulae (6.20), (6.21) for the optimal consumption and portfolio decisions are the same as those of the previous case.

Remark 7.10. Within the Markovian context of nonrandom coefficients, Detemple and Zapatero (1992) obtain a closed-form representation for the optimal portfolio via an application of the Clark (1970) formula; this reduces to “feedback form” for the logarithmic utility function. This feedback formula now becomes a special case of the expression (6.25) (Example 7.9, Case 1) established in Remark 6.10 for an arbitrary utility function.

Remark 7.11. When $\delta(\cdot) = \alpha(\cdot) = 0$ and $z = 0$, i.e., without habit formation, we have $\mu(\cdot) = 1$ from (6.6), so the analysis remains valid for a *random* interest rate process $r(\cdot)$ as well. Thus, this paper generalizes also the dynamic programming/partial differential equation approach to classical utility optimization, developed by Karatzas, Lehoczky, and Shreve (1987) in the special context of deterministic coefficients.

8. Appendix. In this section we provide a stochastic Feynman–Kac formula and an equivalent FBSDE representation for the BSPDEs considered in Lemmata 6.6 and 7.1; our reasoning proceeds along the lines of Ma and Yong (1997).

Preparing the ground of our approach, we state the following implication of the generalized Itô–Kunita–Wentzell formula (see, e.g., Kunita (1990), Section 3.3, pp. 92–93). This enables us to carry out computations in a stochastically modulated dynamic framework; see also Lemma 7.2, Theorem 7.4, and Proposition 8.3.

Proposition 8.1. *Suppose that the random field* $\mathbf{F} : [0,T] \times \mathbb{R}^n \times \Omega \to \mathbb{R}$ *is of class* $C^{0,2}([0,T] \times \mathbb{R}^n)$ *and satisfies*

$$\mathbf{F}(t,\mathrm{x}) = \mathbf{F}(0,\mathrm{x}) + \int_0^t \mathbf{f}(s,\mathrm{x})ds + \int_0^t \mathbf{g}^*(s,\mathrm{x})dW(s) \quad \forall\ (t,\mathrm{x}) \in [0,T] \times \mathbb{R}^n,$$

almost surely. Here $\mathbf{g} = \big(\mathbf{g}^{(1)}, \ldots, \mathbf{g}^{(d)}\big)$, $\mathbf{g}^{(j)} : [0,T] \times \mathbb{R}^n \times \Omega \to \mathbb{R}$, $j = 1, \ldots, d$, *are* $C^{0,2}([0,T] \times \mathbb{R}^n)$, $\mathbb{F}$*-adapted random fields, and* $\mathbf{f} : [0,T] \times \mathbb{R}^n \times \Omega \to \mathbb{R}$ *is a* $C^{0,1}([0,T] \times \mathbb{R}^n)$ *random field. Furthermore, let* $\mathbf{X} = \big(\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)}\big)$ *be a vector of continuous semimartingales with decompositions*

$$\mathbf{X}^{(i)}(t) = \mathbf{X}^{(i)}(0) + \int_0^t \mathbf{b}^{(i)}(s)ds + \int_0^t \big(\mathbf{h}^{(i)}(s)\big)^* dW(s), \quad 0 \le t \le T,$$

for $i = 1, \ldots, n$, *where* $\mathbf{h}^{(i)} = \big(\mathbf{h}^{(i,1)}, \ldots, \mathbf{h}^{(i,d)}\big)$ *is an* $\mathbb{F}$*-progressively measurable, almost surely square integrable vector process, and* $\mathbf{b}^{(i)}(\cdot)$ *is an almost surely integrable process. Then* $\mathbf{F}(\cdot, \mathbf{X}(\cdot))$ *is also a continuous semimartingale, with decomposition*

$$
\begin{aligned}
\mathbf{F}\big(t,\mathbf{X}(t)\big) =& \mathbf{F}\big(0,\mathbf{X}(0)\big) + \int_0^t \mathbf{f}\big(s,\mathbf{X}(s)\big)ds + \int_0^t \mathbf{g}^*\big(s,\mathbf{X}(s)\big)dW(s) \\
&+ \sum_{i=1}^n \int_0^t \frac{\partial}{\partial \mathbf{x}_i}\mathbf{F}\big(s,\mathbf{X}(s)\big)\mathbf{b}^{(i)}(s)ds \\
&+ \sum_{i=1}^n \int_0^t \frac{\partial}{\partial \mathbf{x}_i}\mathbf{F}\big(s,\mathbf{X}(s)\big)\big(\mathbf{h}^{(i)}(s)\big)^* dW(s) \\
&+ \sum_{j=1}^d \sum_{i=1}^n \int_0^t \frac{\partial}{\partial \mathbf{x}_i}\mathbf{g}^{(j)}\big(s,\mathbf{X}(s)\big)\,\mathbf{h}^{(i,j)}(s)\,ds \qquad (8.1) \\
&+ \frac{1}{2}\sum_{\ell=1}^d \sum_{i=1}^n \sum_{k=1}^n \int_0^t \frac{\partial^2}{\partial \mathbf{x}_i \partial \mathbf{x}_k}\mathbf{F}\big(s,\mathbf{X}(s)\big)\,\mathbf{h}^{(i,\ell)}(s)\mathbf{h}^{(k,\ell)}(s)\,ds, \qquad 0 \le t \le T.
\end{aligned}
$$

By analogy with (6.9) and (6.11), for every $(t,y) \in [0,T] \times \mathbb{R}^+$, we regard the stochastic process

$$
\Upsilon^{(t,y)}(s) \triangleq y\,\ell(t)\,\exp\left\{\int_t^s \left[m(v) - \frac{1}{2}\,\|\nu(v)\|^2\right] dv + \int_t^s \nu^*(v)dW(v)\right\} \tag{8.2}
$$

for $t \le s \le T$, which is the unique solution of the linear *forward stochastic differential equation* (FSDE)

$$
d\Upsilon^{(t,y)}(s) = \Upsilon^{(t,y)}(s)\big[m(s)ds + \nu^*(s)dW(s)\big], \quad t < s \le T, \tag{8.3}
$$

$$
\Upsilon^{(t,y)}(t) = y\ell(t); \tag{8.4}
$$

we assume that $m(\cdot)$, $\ell(\cdot)$, and $\nu(\cdot) = (\nu_1(\cdot), \dots, \nu_d(\cdot))^*$ are $\mathbb{F}$-progressively measurable stochastic processes and satisfy the conditions

$$
\int_0^T |m(t)|dt < \infty, \quad \ell(\cdot) \ge 1, \int_0^T \ell(t)dt < \infty, \quad \text{almost surely,}
$$
$$
\text{and} \quad E\int_0^T \|\nu(t)\|^2 dt < \infty.
$$

We define also the random field $\mathfrak{L} : [0,T] \times \mathbb{R}^+ \times \Omega \to \mathbb{R}$ by

$$
\mathfrak{L}(t,y) \triangleq E_t\left[\int_t^T e^{\int_t^s \rho(v)dv} q(s) N\big(s,\Upsilon^{(t,y)}(s)\big)ds\right], \tag{8.5}
$$

for a given jointly continuous function $N : [0,T] \times \mathbb{R}^+ \to \mathbb{R}$, and given $\mathbb{F}$-progressively measurable stochastic processes $\rho(\cdot)$ and $q(\cdot)$, which satisfy the preceding conditions imposed for $m(\cdot)$ and $\ell(\cdot)$, respectively. Notice that the random fields $\mathfrak{X}$ and $\mathfrak{G}$ of (6.13) and (7.7), respectively, now become special cases of $\mathfrak{L}$ for appropriate choices of the probability measure, the processes $m(\cdot)$, $\nu(\cdot)$, $\ell(\cdot)$, $\rho(\cdot)$, $q(\cdot)$, and the function N.

In order to comply with Assumptions 6.1 and 6.2 as well, we shall make the following hypotheses and then state a unifying result for the BSPDEs of Lemmata 6.6 and 7.1.

Assumption 8.2. The processes $m(\cdot)$, $\nu(\cdot)$, $\rho(\cdot)$ are continuous, and $q(\cdot)$, $\ell(\cdot)$ are differentiable. Additionally, for any $t \in [0,T]$, the function $N(t,\cdot)$ is of class $C^4(\mathbb{R}^+)$ and satisfies the polynomial growth condition posited in Remark 6.3 for $u \circ I$.

PROPOSITION 8.3. *Under Assumption* 8.2, *there exists a random field* $\Psi^{\mathfrak{L}} : [0,T] \times \mathbb{R}^+ \times \Omega \to \mathbb{R}^d$ *such that the pair* $(\mathfrak{L}, \Psi^{\mathfrak{L}})$, *where* $\mathfrak{L}$ *is given by* (8.5), *belongs to the class* $C_{\mathbb{F}}\big([0,T];\mathbb{L}^2(\Omega;C^3(\mathbb{R}^+))\big) \times \mathbb{L}^2_{\mathbb{F}}\big(0,T;\mathbb{L}^2(\Omega;C^2(\mathbb{R}^+;\mathbb{R}^d))\big)$ *and is the almost sure unique solution of the Cauchy problem*

$$\begin{aligned}-d\mathfrak{L}(t,y) = \Bigg[&\frac{1}{2}\|\nu(t)\|^2 y^2 \mathfrak{L}_{yy}(t,y) + \left(m(t) - \frac{\ell'(t)}{\ell(t)}\right) y\mathfrak{L}_y(t,y) + \rho(t)\mathfrak{L}(t,y) \\ &+ \nu^*(t) y \Psi^{\mathfrak{L}}_y(t,y) + q(t)N(t, y\ell(t))\Bigg] dt - \big(\Psi^{\mathfrak{L}}(t,y)\big)^* dW(t)\end{aligned} \tag{8.6}$$

on $[0,T) \times \mathbb{R}^+$, *subject to the terminal condition*

$$\mathfrak{L}(T,y) = 0 \quad \textit{on } \mathbb{R}^+. \tag{8.7}$$

Furthermore, for every $(t,y) \in [0,T] \times \mathbb{R}^+$, *consider the stochastic processes*

$$\mathcal{Q}^{(t,y)}(s) \triangleq \mathfrak{L}\left(s, \frac{\Upsilon^{(t,y)}(s)}{\ell(s)}\right), \quad s \in [t,T], \quad \textit{and} \tag{8.8}$$

$$\mathcal{Z}^{(t,y)}(s) \triangleq \nu(s)\frac{\Upsilon^{(t,y)}(s)}{\ell(s)}\mathfrak{L}_y\left(s, \frac{\Upsilon^{(t,y)}(s)}{\ell(s)}\right) + \Psi^{\mathfrak{L}}\left(s, \frac{\Upsilon^{(t,y)}(s)}{\ell(s)}\right), \quad s \in [t,T], \tag{8.9}$$

where $\Upsilon^{(t,y)}(\cdot)$ *satisfies the FSDE of* (8.3) *and* (8.4). *Then the pair* $\big(\mathcal{Q}^{(t,y)}, \mathcal{Z}^{(t,y)}\big)$ *constitutes the unique* $\mathbb{F}$*-adapted solution of the associated linear* backward stochastic differential equation *(BSDE)*

$$\begin{aligned}-d\mathcal{Q}^{(t,y)}(s) = &\Big[\rho(s)\mathcal{Q}^{(t,y)}(s) + q(s)N\big(s, \Upsilon^{(t,y)}(s)\big)\Big] ds \\ &- \big(\mathcal{Z}^{(t,y)}(s)\big)^* dW(s), \quad t \le s < T,\end{aligned} \tag{8.10}$$

$$\mathcal{Q}^{(t,y)}(T) = 0. \tag{8.11}$$

Consequently, the FSDE of (8.3), (8.4) and the BSDE of (8.10), (8.11) constitute an equivalent FBSDE formulation for the BSPDE of (8.6) and (8.7).

Proof of Proposition 8.3. For each $(t,\eta) \in [0,T] \times \mathbb{R}^+$, we define the process

$$R^{(t,\eta)}(s) \triangleq \eta + \log\left(\frac{\Upsilon^{(t,1)}(s)}{\ell(t)}\right), \quad t \le s \le T, \tag{8.12}$$

which, due to (8.3) and (8.4), satisfies the dynamics

$$dR^{(t,\eta)}(s) = \left[m(s) - \frac{1}{2}\|\nu(s)\|^2\right] ds + \big(\nu(s)\big)^* dW(s), \quad t < s \le T, \tag{8.13}$$

$$R^{(t,\eta)}(t) = \eta. \tag{8.14}$$

Moreover, from (8.5) we get the relationship

$$\mathfrak{L}(t,y) = \mathfrak{U}\big(t, \log(y\ell(t))\big) \tag{8.15}$$

for $(t,y) \in [0,T] \times \mathbb{R}^+$, in terms of the random field $\mathfrak{U} : [0,T] \times \mathbb{R}^+ \times \Omega \to \mathbb{R}$ given by

$$\mathfrak{U}(t,\eta) \triangleq E_t \left[\int_t^T e^{\int_t^s \rho(v)dv} q(s) N\big(s, e^{R^{(t,\eta)}(s)}\big) ds \right]. \tag{8.16}$$

According to Assumption 8.2, (8.13), (8.14), and the study of parabolic BSPDEs by Ma and Yong (1997), Corollary 6.2, page 76, there exists a random field $\Psi : [0,T] \times \mathbb{R}^+ \times \Omega \to \mathbb{R}^d$ such that the pair $(\mathfrak{U}, \Psi)$, with $\mathfrak{U}$ given by (8.16), is of class $C_{\mathbb{F}}\big([0,T]; \mathbb{L}^2(\Omega; C^3(\mathbb{R}^+))\big) \times \mathbb{L}^2_{\mathbb{F}}\big(0,T; \mathbb{L}^2(\Omega; C^2(\mathbb{R}^+; \mathbb{R}^d))\big)$ and is, almost surely, the unique solution of the parabolic BSPDE

$$\begin{aligned} -d\mathfrak{U}(t,\eta) = \bigg[& \frac{1}{2}\|\nu(t)\|^2 \mathfrak{U}_{\eta\eta}(t,\eta) + \left(m(t) - \frac{1}{2}\|\nu(t)\|^2 \right) \mathfrak{U}_\eta(t,\eta) \\ & + \rho(t)\mathfrak{U}(t,\eta) + \nu^*(t)\Psi_\eta(t,\eta) + q(t)N(t,e^\eta) \bigg] dt - \big(\Psi(t,\eta)\big)^* dW(t) \end{aligned} \tag{8.17}$$

for $\eta \in \mathbb{R}$, $0 \le t < T$, with terminal condition

$$\mathfrak{U}(T,\eta) = 0, \quad \eta \in \mathbb{R}. \tag{8.18}$$

For any $(t,\eta) \in [0,T] \times \mathbb{R}^+$, Theorem 6.1 of the above citation (page 75) implies also that the processes

$$\mathcal{S}^{(t,\eta)}(s) \triangleq \mathfrak{U}\Big(s, R^{(t,\eta)}(s)\Big), \quad t \le s \le T, \quad \text{and} \tag{8.19}$$

$$\mathcal{H}^{(t,\eta)}(s) \triangleq \nu(s)\mathfrak{U}_\eta\Big(s, R^{(t,\eta)}(s)\Big) + \Psi\Big(s, R^{(t,\eta)}(s)\Big), \quad t \le s \le T, \tag{8.20}$$

form the unique $\mathbb{F}$-adapted solution pair $\big(\mathcal{S}^{(t,\eta)}, \mathcal{H}^{(t,\eta)}\big)$ of the linear BSDE

$$\begin{aligned} -d\mathcal{S}^{(t,\eta)}(s) = & \Big[\rho(s)\mathcal{S}^{(t,\eta)}(s) + q(s)N\big(s, e^{R^{(t,\eta)}(s)}\big) \Big] ds \\ & - \big(\mathcal{H}^{(t,\eta)}(s)\big)^* dW(s), \quad t \le s < T, \end{aligned} \tag{8.21}$$

$$\mathcal{S}^{(t,\eta)}(T) = 0. \tag{8.22}$$

Recalling (8.15), define the random field $\Psi^{\mathfrak{L}} : [0,T] \times \mathbb{R}^+ \times \Omega \to \mathbb{R}^d$ by

$$\Psi^{\mathfrak{L}}(t,y) \triangleq \Psi\big(t, \log(y\ell(t))\big). \tag{8.23}$$

Thanks to (8.15) and (8.23), the pair of random fields $(\mathfrak{L}, \Psi^{\mathfrak{L}})$ has the same regularity as the pair $(\mathfrak{U}, \Psi)$. Thus, the relationships (8.15) and (8.18), and an application of the generalized Itô–Kunita–Wentzell formula (Proposition 8.1), in conjunction with (8.17), yield that $(\mathfrak{L}, \Psi^{\mathfrak{L}})$ is the unique solution of the Cauchy problem (8.6), (8.7). Finally, we couple (8.12) and (8.15) to obtain the equations

$$\mathcal{Q}^{(t,y)}(\cdot) = \mathcal{S}^{(t, \log(y\ell(t)))}(\cdot) \qquad \text{and} \qquad \mathcal{Z}^{(t,y)}(\cdot) = \mathcal{H}^{(t, \log(y\ell(t)))}(\cdot)$$

for $(t,y) \in [0,T] \times \mathbb{R}^+$, and the second part of the proposition follows readily by the $\mathbb{F}$-adaptivity of $\big(\mathcal{S}^{(t,\eta)}, \mathcal{H}^{(t,\eta)}\big)$ and the BSDE of (8.21) and (8.22). □

9. Conclusion. We have studied various aspects of portfolio/consumption optimization in the presence of addictive habits, in complete financial markets. The effective state space for the optimal wealth and standard of living processes was identified as a random wedge, and the investor's value function was shown to have properties similar to those of a utility function. Of particular interest was the interplay between dynamic programming principles and SPDE theory; this led to the characterization of the value random field as a solution of a (highly nonlinear) Hamilton–Jacobi–Bellman BSPDE. The convex dual of the value random field turned out to be the unique solution of a parabolic BSPDE. A by-product of this analysis was an additional representation for the optimal investment-consumption policies on the current level of the optimal wealth and standard of living processes.

The existence of an optimal portfolio/consumption pair in an *incomplete market* (that is, when the number of stocks is strictly smaller than the dimension of the driving Brownian motion) is an open question. Following the duality methodology developed in Karatzas et al. (1991), one can complete the market with fictitious stocks by parametrizing a certain family of exponential local martingales, which includes $Z(\cdot)$ of (2.6) and gives rise to an analogous class of state-price density processes. An associated dual optimization problem can be defined in terms of the respective parametrized "adjusted" state-price density processes, such that a possible minimizer induces a null demand for the fictitious stocks. But in the context of habit formation, the dual functional fails to be convex with respect to the dual parameter; thus, new methodologies will most likely have to be developed to handle the problem. We leave this issue as an open question for future research.

Acknowledgments. We are deeply indebted to the referees for their very careful reading of the paper and for their many suggestions that improved it very considerably. We are also grateful to Professors Jin Ma, Daniel Ocone, and Peter Bank for their valuable comments.

REFERENCES

P. Bank and H. Föllmer (2003), *American Options. Multi-armed Bandits, and Optimal Consumption Plans: A Unified View*, Lecture Notes in Math. 1814, Springer-Verlag, New York, pp. 1–42.

P. Bank and F. Riedel (2000), *Non-time-additive utility optimization—the case of certainty*, J. Math. Econom., 33, pp. 271–290.

P. Bank and F. Riedel (2001), *Optimal consumption choice with intertemporal substitution*, Ann. Appl. Probab., 11, pp. 750–788.

J. M. Bismut (1973), *Conjugate convex functions in optimal stochastic control*, J. Math. Anal. Appl., 44, pp. 384–404.

R. Buckdahn and J. Ma (2001a), *Stochastic viscosity solutions for nonlinear stochastic partial differential equations. Part* I, Stochastic Process. Appl., 93, pp. 181–204.

R. Buckdahn and J. Ma (2001b), *Stochastic viscosity solutions for nonlinear stochastic partial differential equations. Part* II, Stochastic Process. Appl., 93, pp. 205–228.

R. Buckdahn and J. Ma (2007), *Pathwise stochastic control problems and stochastic HJB equations*, SIAM J. Control Optim., 45, pp. 2224–2256.

D. A. Chapman (1998), *Habit formation and aggregate consumption*, Econometrica, 66, pp. 1223–1230.

J. M. C. Clark (1970), *The representation of functionals of Brownian motion at stochastic integrals*, Ann. Math. Statist., 41, pp. 1282–1295.

G. M. Constantinides (1990), *Habit formation: A resolution of the equity premium puzzle*, J. Polit. Econ., 98, pp. 519–543.

J. Cox and C. F. Huang (1989), *Optimal consumption and portfolio policies when asset prices follow a diffusion process*, J. Econom. Theory, 49, pp. 33–83.

J. B. Detemple and I. Karatzas (2003), *Non addictive habits: Optimal portfolio-consumption policies*, J. Econom. Theory, 113, pp. 265–285.

J. B. Detemple and F. Zapatero (1991), *Asset prices in an exchange economy with habit-formation*, Econometrica, 59, pp. 1633–1657.

J. B. Detemple and F. Zapatero (1992), *Optimal consumption-portfolio policies with habit formation*, Math. Finance, 2, pp. 251–274.

J. Heaton (1993), *The interaction between time-nonseparable preferences and time-aggregation*, Econometrica, 61, pp. 353–385.

A. Hindy and C. F. Huang (1993), *Optimal portfolio and consumption rules with durability and local substitution*, Econometrica, 61, pp. 85–121.

A. Hindy, C. F. Huang, and D. Kreps (1992), *On intertemporal preferences in continuous time: The case of certainty*, J. Math. Econom., 21, pp. 401–440.

I. Karatzas (1989), *Optimization problems in theory of continuous trading*, SIAM J. Control Optim., 27, pp. 1221–1259.

I. Karatzas, J. P. Lehoczky, S. P. Sethi, and S. E. Shreve (1986), *Explicit solution of a general consumption/investment problem*, Math. Oper. Res., 11, pp. 261–294.

I. Karatzas, J. P. Lehoczky, and S. E. Shreve (1987), *Optimal portfolio and consumption decisions for a "small investor" on a finite horizon*, SIAM J. Control Optim., 25, pp. 1557–1586.

I. Karatzas, J. P. Lehoczky, S. E. Shreve, and G.-L. Xu (1991), *Martingale and duality methods for utility maximization in an incomplete market*, SIAM J. Control Optim., 29, pp. 702–730.

I. Karatzas and S. E. Shreve (1991), *Brownian Motion and Stochastic Calculus*, 2nd ed., Springer-Verlag, New York.

I. Karatzas and S. E. Shreve (1998), *Methods of Mathematical Finance*, Springer-Verlag, New York.

H. Kunita (1990), *Stochastic Flows and Stochastic Differential Equations*, Cambridge Stud. Adv. Math. 24, Cambridge University Press, Cambridge, UK.

P.-L. Lions and P. E. Souganidis (1998a), *Fully nonlinear stochastic partial differential equations*, C. R. Acad. Sci. Paris Sér. 1 Math., 326, pp. 1085–1092.

P.-L. Lions and P. E. Souganidis (1998b), *Fully nonlinear stochastic partial differential equations: Non-smooth equations and applications*, C. R. Acad. Sci. Paris Sér. 1 Math., 327, pp. 735–741.

J. Ma and J. Yong (1997), *Adapted solution of a degenerate backward SPDE, with applications*, Stochastic Process. Appl., 70, pp. 59–84.

J. Ma and J. Yong (1998), *Forward-Backward Stochastic Differential Equations and Their Applications*, Lecture Notes in Math. 1702, Springer-Verlag, New York.

J. Ma and J. Yong (1999), *On linear, degenerate backward stochastic partial differential equations*, Probab. Theory Related Fields, 113, pp. 135–170.

R. C. Merton (1969), *Lifetime portfolio selection under uncertainty: The continuous-time case*, Rev. Econom. Statist., 51, pp. 247–257.

R. C. Merton (1971), *Optimum consumption and portfolio rules in a continuous time model*, J. Econom. Theory, 3, pp. 373–413.

D. L. Ocone and I. Karatzas (1991), *A generalized Clark representation formula, with application to optimal portfolios*, Stochastics, 34, pp. 187–220.

S. Peng (1992), *Stochastic Hamilton–Jacobi–Bellman equations*, SIAM J. Control Optim., 30, pp. 284–304.

S. R. Pliska (1986), *A stochastic calculus model of continuous trading: Optimal portfolio*, Math. Oper. Res., 11, pp. 371–382.

R. T. Rockafellar (1970), *Convex Analysis*, Princeton University Press, Princeton, NJ.

M. Schroder and C. Skiadas (2002), *An isomorphism between asset-pricing with and without habit-formation*, Rev. Financial Stud., 15, pp. 1189–1221.

S. M. Sundaresan (1989), *Intertemporally dependent preferences and volatility of consumption and wealth*, Rev. Financial Stud., 2, pp. 73–89.

G. L. Xu (1990), *A Duality Method for Optimal Consumption and Investment under Short-Selling Prohibition*, Doctoral dissertation, Department of Mathematics, Carnegie Mellon University, Pittsburgh, PA.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 2, pp. 521–550

© 2009 Society for Industrial and Applied Mathematics

ANALYSIS OF THE HUM CONTROL OPERATOR AND EXACT CONTROLLABILITY FOR SEMILINEAR WAVES IN UNIFORM TIME[*]

B. DEHMAN[†] AND G. LEBEAU[‡]

Abstract. We present a regularity result for the HUM optimal control associated with the interior control of linear waves. We use this analysis, together with Strichartz inequalities, to get results on the exact controllability for subcritical nonlinear waves in a bounded domain of $\mathbb{R}^3$.

Key words. HUM control operator, semilinear wave equation

AMS subject classifications. 35L20, 35L70, 74J30, 93B05, 93B07, 93C20

DOI. 10.1137/070712067

1. Introduction and statement of the results.

The main goal of this paper is to study the exact controllability of subcritical semilinear waves in bounded domains of $\mathbb{R}^3$ with Dirichlet boundary condition. We are able to show, using Littlewood–Paley theory and the Strichartz inequalities obtained in [BLP07], that the control problem can be understood frequency by frequency; in other words, the energy of each scale of the control function depends (almost) only on the energy of the same scale in the states one wants to control. This is made precise in Theorem 1.8 and its proof. On the way, we observe that high frequencies (small scales) are entirely governed by the linear part of the equation, a fact which comes clearly from the subcritical nature of the problem. On the other hand, it turns out that low frequencies (large scales) are much more difficult to analyze, since they clearly remain associated with a nonlinear dynamic, and, for this reason, in Theorem 1.8, we will still assume a smallness condition on the energy of the low frequency part of the states we want to control.

In order to do this reduction to low frequency, we prove in Theorem 1.4 that the optimal L^2 control operator Λ associated with the linear wave equation almost commutes with frequency localization. Moreover, in Theorem 1.3, we will prove that Λ preserves all of the Sobolev spaces. (As an easy fact, we will also show in Theorem 4.1 that, in the case of a manifold without boundary, Λ is indeed a pseudodifferential operator.)

The paper is organized as follows. We first recall some basic facts on optimal linear control theory and the so-called HUM method. Then we recall the geometric control condition of [BLR92]. We then state our main results in Theorems 1.3, 1.4, and 1.8. In section 2, we prove Theorems 1.3 and 1.4, and, in section 3, we prove Theorem 1.8. Finally, in the appendix, we first prove Theorem 4.1 on the properties of the optimal L^2 control operator on a manifold without boundary, and we prove some useful lemmas on the action of a smooth multiplier in the case of bounded

[*] Received by the editors December 31, 2007; accepted for publication (in revised form) September 10, 2008; published electronically February 13, 2009.
http://www.siam.org/journals/sicon/48-2/71206.html

[†] Département de Mathématiques, Faculté des sciences de Tunis et Enit-Lamsin, 2092 El Manar, Tunisie (Belhassen.Dehman@fst.rnu.tn). This author's research was supported by the Tunisian Ministry for Scientific Research and Technology within the LAB-STI 02 program.

[‡] Département de Mathématiques, Université de Nice Sophia-Antipolis, Parc Valrose, 06108 Nice Cedex 02, France (lebeau@math.unice.fr).

domains of $\mathbb{R}^d$; then we recall what we need on Strichartz inequalities, and we prove the composition theorem (Theorem 4.7), which is one of the ingredients of the proof of Theorem 1.8.

The interested reader will find surprisingly good illustrations of the theoretical results of this paper on the analytic structure of the optimal control operator in [LN08].

1.1. The HUM method. The problem of controllability for linear evolution equations and systems has a long history (see the review of Russell in [Rus78]). Here we recall briefly the so-called HUM method, for which we refer the reader to the book [Lio88] of Lions, since we shall use in what follows some basic notation and facts in controllability theory.

Let H be a Hilbert space, and let $U(t) = e^{itA},\ t \geq 0$, be a continuous semigroup of contractions on H, with generator A. Let B be a bounded operator on H. Then, for any $g(t) \in L^1([0, \infty[, H)$, the evolution equation

$$(\partial_t - iA)f = Bg, \quad f(0) = 0 \tag{1.1}$$

admits a unique solution $f = S(g) \in C^0([0, \infty[, H)$ given by the Duhamel formula

$$f(t) = \int_0^t e^{i(t-s)A} Bg(s)ds. \tag{1.2}$$

Let $T > 0$ be given. Let $\mathcal{R}_T$ be the reachable set at time T:

$$\mathcal{R}_T = \{f \in H,\ \exists g \in L^2([0, T], H),\ f = S(g)(T)\}. \tag{1.3}$$

Then $\mathcal{R}_T$ is a linear subspace of H, and is the set of states of the system that one can reach in time T, starting at rest, with the action of an L^2 source g filtered by the control operator B. The control problem is to give an accurate description of $\mathcal{R}_T$, and exact controllability is equivalent to the equality $\mathcal{R}_T = H$. Let us recall some basic facts.

Let $\mathcal{H} = L^2([0, T], H)$. Let $\mathcal{F}$ be the closed subspace of $\mathcal{H}$ spanned by solutions of the adjoint evolution equation

$$\mathcal{F} = \{h \in \mathcal{H},\ (\partial_t - iA^*)h = 0,\ h(T) = h_T \in H\}. \tag{1.4}$$

Let $\mathcal{B}^*$ be the adjoint of the operator $g \mapsto S(g)(T)$. Then $\mathcal{B}^*$ is the bounded operator from H into $\mathcal{H}$, with values in $B^*\mathcal{F}$:

$$\mathcal{B}^*(h_T)(t) = B^* e^{-i(T-t)A^*} h_T. \tag{1.5}$$

For any $g \in L^2([0, T], H)$, one has, with $f_T = S(g)(T)$ and $h(s) = e^{-i(T-s)A^*} h_T$, the fundamental identity

$$(f_T|h_T)_H = \int_0^T (Bg(s)|h(s))ds = (g|\mathcal{B}^*(h_T))_{\mathcal{H}}. \tag{1.6}$$

From (1.6), one gets easily that the following holds true:

$$\mathcal{R}_T \text{ is a dense subspace of } H \iff \mathcal{B}^* \text{ is an injective operator,} \tag{1.7}$$

which shows that approximate controllability is equivalent to a uniqueness result on the adjoint equation. Moreover, one gets from (1.6), using the Riesz and closed graph theorems, that the following holds true:

$$e^{iTA}H \subset \mathcal{R}_T \iff \exists C, \quad \|e^{-iTA^*}h\|_H \le C\|\mathcal{B}^*h\|_{\mathcal{H}} \quad \forall h \in H, \tag{1.8}$$

$$\mathcal{R}_T = H \iff \exists C, \quad \|h\|_H \le C\|\mathcal{B}^*h\|_{\mathcal{H}} \quad \forall h \in H. \tag{1.9}$$

Both (1.8) and (1.9) are observability inequalities, and $\mathcal{B}^*$ is called the observability operator. The observability inequality (1.8) is used in the study of parabolic equations like the heat equation. Since here we will work with wave equations, which are reversible in time, $U(t) = e^{itA}$ will be a group of isometries, or at least a well-defined group of isomorphisms, for all $t \in \mathbb{R}$. In that case, (1.8) and (1.9) are equivalent, and, from now on, we assume that $D(A^*) = D(A)$ and $A^* - A$ is a bounded operator on H. We rewrite the observability inequality (1.9) in the more explicit form

$$\exists C, \quad \|h\|_H^2 \le C\int_0^T \|B^*e^{-isA^*}h\|_H^2 ds \quad \forall h \in H. \tag{1.10}$$

Assuming that (1.10) holds true, then $\mathcal{R}_T = H$, $\mathrm{Im}(\mathcal{B}^*)$ is a closed subspace of $\mathcal{H}$, and $\mathcal{B}^*$ is an isomorphism of H onto $\mathrm{Im}(\mathcal{B}^*)$. For any $f \in H$, let $\mathcal{C}_f$ be the set of control functions g driving 0 to f in time T:

$$\mathcal{C}_f = \left\{g \in L^2([0,T], H), \quad f = \int_0^T e^{i(T-s)A}Bg(s)ds\right\}. \tag{1.11}$$

From (1.6), one gets

$$\mathcal{C}_f = g_0 + (\mathrm{Im}\,\mathcal{B}^*)^{\perp}, \quad g_0 \in \mathrm{Im}\,\mathcal{B}^* \cap \mathcal{C}_f, \tag{1.12}$$

and $g_0 = \mathcal{B}^*h_T$ is the optimal control in the sense that

$$\min\{\|g\|_{L^2([0,T],H)},\ g \in \mathcal{C}_f\} \quad \text{is achieved at } g = g_0. \tag{1.13}$$

Let $\Lambda : H \to H$, $\Lambda(f) = h_T$ be the control map, so that the optimal control g_0 is equal to $g_0(t) = B^*e^{-i(T-t)A^*}\Lambda(f)$. Then Λ is exactly the inverse of the map $M_T : H \to H$ with

$$\begin{aligned} M_T &= \int_0^T m(T-t)dt = \int_0^T m(s)ds, \\ m(s) &= e^{isA}BB^*e^{-isA^*}. \end{aligned} \tag{1.14}$$

Then $m(s) = m^*(s)$ is a bounded, self-adjoint, nonnegative operator on H, and from (1.10) exact controllability is equivalent to

$$\exists C > 0, \quad M_T = \int_0^T m(s)ds \ge C\,\mathrm{Id}. \tag{1.15}$$

The above discussion applies as well when the control operator $B = B(t)$ in (1.1) is a bounded family in $t \in [0,T]$ of bounded operators on H. In that case, exact controllability is still equivalent to (1.15) with M_T given by

$$M_T = \int_0^T e^{i(T-t)A}B(t)B^*(t)e^{-i(T-t)A^*}dt. \tag{1.16}$$

In what follows, we will use a time dependent control operator B (see (1.19)).

1.2. Geometry and the (GCC) assumption. Let M be either a compact, connected, Riemannian manifold of dimension d without boundary or a bounded, connected, and open subset Ω of $\mathbb{R}^d$ with smooth boundary. In the first case, Δ will denote the Laplace–Beltrami operator on M and, if $M = \Omega$, then Δ will denote the usual Laplace operator on $\Omega \subset \mathbb{R}^d$ with the Dirichlet boundary condition on $\partial\Omega$; all of our statements will be true in the more general case where Ω is a relatively compact open regular subset of a Riemannian manifold.

We denote by (ω_j^2) the sequence of eigenvalues of $-\Delta$ and (e_j) the orthonormal basis of $L^2(M)$ constituted by the eigenvectors associated with (ω_j^2):

$$-\Delta e_j = \omega_j^2 e_j, \qquad \|e_j\|_{L^2(M)} = 1.$$

We will have $j \geq 1$ and $\omega_1 > 0$ in the case $M = \Omega$, and $j \geq 0$ and $\omega_0 = 0, e_0 = (vol(M))^{-1/2}$ in the case M compact. For any $s \in \mathbb{R}$, we denote by $H^s(M)$ the usual Sobolev space and by $H^s(M, \Delta)$ the Hilbert space defined by

$$H^s(M, \Delta) = \Big\{ u = \sum_j a_j e_j, \quad \sum_j (1 + \omega_j^2)^s \, |a_j|^2 < \infty \Big\}. \tag{1.17}$$

When M is compact, $H^s(M, \Delta) = H^s(M)$ is the usual Sobolev space. When $M = \Omega$, $H^s(\Omega, \Delta)$ is the domain of $(1 - \Delta)^{s/2}$, and we hope that this notation will not be confusing for the reader; one has in that case $H^1(\Omega, \Delta) = H_0^1(\Omega)$. For $s \geq 0$, $H^s(\Omega, \Delta)$ is a subset of $H^s(\Omega)$, and one has $H^s(\Omega, \Delta) = H^s(\Omega)$ for $0 \leq s < 1/2$, $H^s(\Omega, \Delta) = \{u \in H^s(\Omega),\ u|_{\partial\Omega} = 0\}$ for $1/2 < s < 5/2$, $H^s(\Omega, \Delta) = \{u \in H^s(\Omega),\ u|_{\partial\Omega} = \Delta u|_{\partial\Omega} = 0\}$ for $5/2 < s < 9/2$, and so on. We will use the self-adjoint operator $\lambda = \lambda(x, D_x) = \sqrt{|\Delta|}$. One has $\lambda^2 + \Delta = 0$ and

$$\lambda(x, D_x) \sum_j a_j e_j = \sum_j \omega_j a_j e_j. \tag{1.18}$$

Moreover, if M is compact, then λ is a first order pseudodifferential operator on M.

On the other hand, we will deal in the whole work with a time dependent control operator. More precisely, let $T > 0$ and define

$$\chi(t, x) = \psi(t) \chi_0(x), \tag{1.19}$$

where χ_0 is a real C^∞ function on $\overline{M}$, $\psi \in C^\infty([0, T])$ is flat at $t = 0, T$, and $\psi(t) > 0$ on $]0, T[$. We will also denote by χ_0 the function on $T^*\overline{M}$ defined by $\chi_0(x, \xi) = \chi_0(x)$.

Our controls vectors (see, for instance, system (4.1) or (1.21) below) will be of the form $\chi(t, x)v$, instead of the usual form $\chi(x)v$. As we will see, with this slight modification, the optimal control operator Λ is simpler and we get better results. In particular, in the case of a compact manifold without boundary, we will show in section 4.1 that Λ is a pseudodifferential operator. Let ω be the open subset of M:

$$\omega = \{x \in M, \chi_0(x) \neq 0\}. \tag{1.20}$$

We will always assume that the open set ω of M satisfies the geometric control condition (GCC) of [BLR92] at time T.

(GCC) Every geodesic ray of M travelling with speed 1 and starting at $t = 0$ enters the open set ω in a time $t < T$.

Of course, for $M = \Omega$, these geodesics have to be understood as the projection onto the basis Ω of the generalized bicharacteristic rays of the wave operator, the so-called Melrose–Sjöstrand flow, for which we refer the reader to [Hör85]. We shall denote by $s \to (\gamma_{(x,\xi)}(s), t - s\tau, \tau)$, $s \in \mathbb{R}$, the generalized bicharacteristic ray of the wave operator, issued from (x, ξ, t, τ). In the case $M = \Omega$, we will always assume that there is no contact of infinite order between the boundary $\partial\Omega$ and the bicharacteristic rays of the wave operator in the free space, so that $\gamma_{(x,\xi)}(s)$ is well defined.

1.3. Controllability of linear waves. As we have said, in this work, M will be either a compact, connected, Riemannian manifold of dimension d without boundary or a bounded, connected, and open subset Ω of $\mathbb{R}^d$. In order to make the paper as clear as possible, we have chosen to present as the main result the case of open subsets Ω of $\mathbb{R}^d$. The corresponding results in the case of a compact manifold without boundary are simpler, and they will be stated and proved in section 4.1.

In the framework of the wave equation in a bounded regular open subset Ω of $\mathbb{R}^d$ with boundary Dirichlet condition, and, for internal control, the problem of controllability is stated in the following way. Let T be a positive time and $\chi(t, x)$ be as in (1.19). For a given $(u_0, u_1) \in H_0^1(\Omega) \times L^2(\Omega)$, the problem is to find a source $v(t, x) \in L^2(0, T; L^2(\Omega))$ such that the solution of the system

$$(1.21) \qquad \begin{cases} \Box u = \chi v \quad \text{in }]0, +\infty[\times\Omega, \\ u_{|\partial\Omega} = 0, \quad t > 0, \\ (u|_{t=0}, \partial_t u|_{t=0}) = (0, 0) \end{cases}$$

reaches the state $(u(T), D_t u(T)) = (u_0, u_1)$ at time T. The HUM method consists in taking the control function v in (1.21) in the form $v = \chi w$, where w is a solution of the dual problem

$$(1.22) \qquad \begin{cases} \Box w = 0 \quad \text{in }]0, +\infty[\times\Omega, \\ w_{|\partial\Omega} = 0, \quad t > 0, \\ (w_0, w_1) = \underline{w}_0 \in E_{-1} = L^2(\Omega) \times H^{-1}(\Omega). \end{cases}$$

By the above discussion, exact controllability is equivalent to the invertibility of the operator M_T given by (1.16), and the optimal control map Λ is given by $\Lambda^{-1} = M_T$. In order to compute the operator M_T, let us first make the obvious algebraic reduction of the wave equation to a first order system like (1.1). Let A be the matrix

$$(1.23) \qquad iA = \begin{pmatrix} 0 & \mathrm{Id} \\ \triangle & 0 \end{pmatrix}.$$

Then A is an unbounded self-adjoint operator on $H = H_0^1(\Omega) \times L^2(\Omega)$, where the scalar product on $H_0^1(\Omega)$ is $\int_\Omega \nabla u \overline{\nabla v} dx$ and $D(A) = \{\underline{u} \in H,\ A(\underline{u}) \in H,\ u_0|_{\partial\Omega} = 0\}$. Set as in (1.18)

$$(1.24) \qquad \lambda = \sqrt{-\triangle_D},$$

where $-\triangle_D$ is the canonical isomorphism from $H_0^1(\Omega)$ onto $H^{-1}(\Omega)$. Then λ is an isomorphism from $H_0^1(\Omega)$ onto $L^2(\Omega)$. The operator $B(t)$ given by

$$(1.25) \qquad B(t) = \begin{pmatrix} 0 & 0 \\ \chi(t, .)\lambda & 0 \end{pmatrix}$$

is bounded on H, and one has

$$B^*(t) = \begin{pmatrix} 0 & \lambda^{-1}\chi(t,.) \\ 0 & 0 \end{pmatrix}. \tag{1.26}$$

The system (1.21) is then equivalent to (1.1), with $f = (u, \partial_t u)$, $g = (\lambda^{-1}v, 0)$, and we observe that the optimal control $g_0(t) = B^* e^{-i(T-t)A^*}\Lambda(f)$ is of the form $g_0 = (\lambda^{-1}\chi w, 0)$, where $w \in L^2([0,T], L^2(\Omega))$ is a solution of (1.22). Thus the exact controllability condition (1.15) is exactly

$$\exists C > 0, \quad M_T = \int_0^T e^{i(T-t)A} \begin{pmatrix} 0 & 0 \\ 0 & \chi^2(t,.) \end{pmatrix} e^{-i(T-t)A} dt \geq C\,\mathrm{Id}. \tag{1.27}$$

Next, we recall the theorem of [BLR92].

Theorem 1.1. *If ω and T are such that* (GCC) *holds true, then M_T is an isomorphism.*

In the early literature, exact controllability was often reached by means of multiplier methods, under the Γ-condition of Lions. Then in [BLR92], the authors showed that for the wave equation, exact controllability (with stability with respect to small perturbations of ω, T) is equivalent to (GCC). This is a microlocal condition (i.e., a property in the cotangent bundle T^*M), linking the couple (ω, T) and the bicharacteristic rays of the wave operator. This then offered the possibility of establishing it through microlocal tools, namely, the propagation of wave front sets or microlocal defect measures (see [Gér91], [BG97], [Leb92]). In this paper, we answer two basic questions.

(1) *Regularity*: If the target state $\underline{u}_T$ belongs to $H^{s+1}(\Omega, \Delta) \times H^s(\Omega, \Delta), s \geq 0$, then in which space does the optimal control $\Lambda(\underline{u}_T)$ live? In other words, is the regularity of the optimal control adjusted to the one of the data to be controlled? When $s > 0$, the solution u of (1.21) attached to the HUM control is, à priori, in $C(0,T; H^1(\Omega, \Delta))$, while we may expect it to be more regular. Hence, the process seems to introduce a loss of smoothness, and this is not very satisfactory.

Remark 1.2. Notice that in [BLR92] the authors prove, for a boundary control problem, observation estimates in each Sobolev space H^s and the existence of a control vector which satisfies the right regularity. This does not give an answer to the question of the regularity of the optimal L^2-control. Also we will see that for an interior control problem the multiplier χ introduces extra difficulties at the boundary.

As already observed in [BLR92], in the boundary case, the right scales of spaces concerning the regularity are the $H^s(\Omega, \Delta)$ since they include the compatibility conditions. The scale $H^s(\Omega)$ is inappropriate. For example, in the case $\Omega =]0, \pi[$ and $\omega =]a, \pi - a[$ with $a > 0$, exact controllability holds true for $T > 2a$. For $\underline{u}_T = (1, 0) \in H^\infty(\Omega) \oplus H^\infty(\Omega)$, any control function v in (1.21) must be singular, since if v is smooth, then we will have $u(T, x) \in H^\infty(\Omega, \Delta)$ and in particular $u(T, 0) = 0$.

(2) *Spectral analysis*: Now, if we assume that the state $\underline{u}_T$ is spectrally supported in some dyadic set $a2^k \leq \omega_j \leq b2^k$, then how are the frequencies of the HUM control $\Lambda(\underline{u}_T)$? Are they also almost localized in the same set if k is large? For instance, if $\underline{u}_T$ has only low frequencies, how are the high frequencies of $\Lambda(\underline{u}_T)$?

To conclude this discussion, we observe that the fact that the HUM control $\Lambda(\underline{u}_T)$ is the solution of a natural variational problem would suggest that the answer to these questions could be affirmative.

In order to separate waves with positive and negative frequencies, we introduce the following obvious splitting. For $\underline{g} = (g_0, g_1) \in H = H_0^1(\Omega) \times L^2(\Omega)$, we set

$$\begin{aligned} g_0 &= \lambda^{-1}(h_+ + h_-), \\ g_1 &= i(h_+ - h_-). \end{aligned} \tag{1.28}$$

One has $h_\pm \in L^2(\Omega)$, and (1.28) gives an identification between H and $L^2(\Omega) \oplus L^2(\Omega)$. Set $\mathcal{U}_\pm(t) = e^{\pm it\lambda}$. Then the operators $\mathcal{U}_\pm(t)$ are isometries on each Sobolev space $H^s(\Omega, \Delta)$, with inverses $\mathcal{U}_\pm(-t)$, and, with this identification, e^{itA} is equal to the diagonal operator

$$e^{itA} = (\mathcal{U}_+(t), \mathcal{U}_-(t)). \tag{1.29}$$

Moreover, using (1.25), we get in the splitting (1.28) that the control operator B is given by

$$B = \frac{1}{2i} \begin{pmatrix} \chi & \chi \\ -\chi & -\chi \end{pmatrix}, \tag{1.30}$$

and we recall that the optimal control map Λ is equal to the inverse of the map M_T:

$$M_T = \int_0^T m(s) ds, \quad m(s) = e^{isA} B(T-s) B^*(T-s) e^{-isA^*}. \tag{1.31}$$

Hence, we obtain from (1.29), (1.30), with the notation $(\chi^2) = \chi^2(T-s, x)$,

$$m(s) = \frac{1}{2} \begin{pmatrix} \mathcal{U}_+(s)(\chi^2)\mathcal{U}_+(-s) & -\mathcal{U}_+(s)(\chi^2)\mathcal{U}_-(-s) \\ -\mathcal{U}_-(s)(\chi^2)\mathcal{U}_+(-s) & \mathcal{U}_-(s)(\chi^2)\mathcal{U}_-(-s) \end{pmatrix}. \tag{1.32}$$

Let $Q_\pm$ be the operators

$$Q_\pm = \int_0^T \mathcal{U}_\pm(s)(\chi^2)\mathcal{U}_\pm(-s) ds, \tag{1.33}$$

and let $\mathcal{T}$ be the operator

$$\mathcal{T} = \int_0^T \mathcal{U}_+(s)(\chi^2)\mathcal{U}_+(s) ds. \tag{1.34}$$

The following theorem gives the algebraic structure of the optimal control map Λ and answers the question of regularity. We shall say that an operator R is smoothing if it maps $L^2(\Omega)$ into $H^\sigma(\Omega, \Delta)$ for all $\sigma \geq 0$.

THEOREM 1.3. *Under* (GCC)*, the operators $Q_\pm$ are isomorphisms on each Sobolev space $H^s(\Omega, \Delta)$ for all $s \geq 0$. The operator $\mathcal{T}$ and its adjoint $\mathcal{T}^*$ are smoothing. Let $L_\pm$ be the inverses of the operators $Q_\pm$. In the splitting* (1.28)*, the HUM control operator Λ is equal to*

$$\Lambda = \begin{pmatrix} 2L_+ & 0 \\ 0 & 2L_- \end{pmatrix} + R, \tag{1.35}$$

where the operator R is smoothing. In particular, Λ is an isomorphism of $H^s(\Omega, \Delta) \oplus H^s(\Omega, \Delta)$ for all $s \geq 0$.

Now we introduce the material needed for the Littlewood–Paley decomposition. Let $\phi \in C^\infty([0,\infty[)$, with $\phi(x) = 1$ for $|x| \leq 1/2$ and $\phi(x) = 0$ for $|x| \geq 1$. Set $\psi(x) = \phi(x) - \phi(2x)$. Then $\psi \in C_0^\infty(\mathbb{R}^*)$, ψ vanishes outside $[1/4, 1]$, and one has

$$\phi(s) + \sum_{k=1}^{\infty} \psi(2^{-k}s) = 1 \qquad \forall s \in [0, \infty[.$$

Set $\psi_0(s) = \phi(s)$ and $\psi_k(s) = \psi(2^{-k}s)$ for $k \geq 1$. We then define the spectral localization operators $\psi_k(D)$, $k \in \mathbb{N}$, in the following way: for $u = \sum_j a_j e_j$, we define

$$\psi_k(D)u = \sum_j \psi_k(\omega_j) a_j e_j. \tag{1.36}$$

One has $\sum_k \psi_k(D) = \mathrm{Id}$ and $\psi_i(D)\psi_j(D) = 0$ for $|i-j| \geq 2$. In addition, we introduce

$$S_k(D) = \sum_{j=0}^{k} \psi_j(D) = \psi_0(2^{-k}D), \qquad k \geq 0. \tag{1.37}$$

Our spectral localization result reads as follows.

THEOREM 1.4. *With the previous notation and under* (GCC)*, there exists $C > 0$ such that for every $k \in \mathbb{N}$ the following inequality holds true:*

$$\begin{aligned} \|\psi_k(D)\Lambda - \Lambda\psi_k(D)\|_{L^2} &\leq C2^{-k}, \\ \|S_k(D)\Lambda - \Lambda S_k(D)\|_{L^2} &\leq C2^{-k}. \end{aligned} \tag{1.38}$$

Remark 1.5. What Theorem 1.4 states is that the HUM control operator Λ, up to a lower order term, acts individually on each frequency block of the solution. For instance, if e_n is the nth vector of the $L^2(\Omega)$-orthonormal basis of eigenvectors of $\triangle$ and if one drives the data (e_n, e_n) to (e_{n+1}, e_{n+1}), then the solution will essentially live at frequency ω_n for n large.

Remark 1.6. The estimates (1.38) still remain valid if we replace the d'Alembertian $\Box$ by the $\Box +$ first order self-adjoint operator; in particular, they are valid for $\Box + c$, $c \in \mathbb{R}$.

1.4. Controllability of subcritical nonlinear waves. In this section, we state in Theorem 1.8 our result concerning the exact controllability of the subcritical semilinear wave equation, which constitutes the second main goal of this paper. All of the functions here will take real values. We still denote by Ω an open and regular subset of $\mathbb{R}^3$. Let f be a function from $\mathbb{R}$ to $\mathbb{R}$, of class C^3, satisfying the following conditions:

$$f(0) = 0, \quad sf(s) \geq 0 \qquad \forall s \in \mathbb{R}, \tag{1.39}$$

$$\left| f^{(j)}(s) \right| \leq C(1 + |s|)^{p-j} \qquad \text{for} \quad j = 1, 2, 3 \tag{1.40}$$

with $C > 0$ and p a real number such that $1 \leq p < 5$. We first recall the theorem of [BLP07] on the well posedness of the semilinear wave equation with boundary Dirichlet condition in Ω.

THEOREM 1.7. *For every $T > 0$, $E_0 > 0$, and $g \in L^1(]0, T[, L^2(\Omega))$ the semilinear and defocusing wave equation*

$$\left\{ \begin{array}{l} \Box u + f(u) = g \qquad in \quad]0, +\infty[\times\Omega, \qquad u_{|\partial\Omega} = 0, \\ \qquad\qquad \|(u|_{t=0}, \partial_t u|_{t=0})\|_{H_0^1 \times L^2} \leq E_0 \end{array} \right. \tag{1.41}$$

admits in the space $C^0(0,T;H_0^1(\Omega))\cap C^1(0,T;L^2(\Omega))$ *a unique solution* u *satisfying*

$$\|u\|_{L^5(0,T;W_0^{\frac{3}{10},5}(\Omega))}\le C \tag{1.42}$$

for some positive constant $C=C(T,E_0,\|g\|_{L^1(0,T;L^2(\Omega))})$*. In particular, for every* $q\in[5,+\infty]$*, and* r *with* $1/q+1/r=1/2$*, there exists a positive constant* $C'=C'(T,q,E_0,\|g\|_{L^1(0,T;L^2(\Omega))})$ *such that*

$$\|u\|_{L^q(0,T;L^{3r}(\Omega))}\le C'. \tag{1.43}$$

With $F(u)=\int_0^u f(s)ds\ge 0$, the nonlinear energy of the solution u of (1.41) is equal to

$$E(u)(t)=\frac{1}{2}\int_M((\partial_t u(t))^2+|\nabla_x u(t)|^2)dx+\int_M F(u(t))dx. \tag{1.44}$$

Our result reads as follows.

THEOREM 1.8. *Assume that* (GCC) *holds true. For any given* $E_0>0$*, there exist* $r>0$ *small and* $k\in\mathbb{N}$ *large, such that for every initial and final data* $\underline{u}_0$*,* $\underline{u}_1$ *in* $H_0^1(\Omega)\times L^2(\Omega)$ *satisfying*

$$\begin{aligned} &\|\underline{u}_0\|_{H_0^1\times L^2}\le E_0, && \|\underline{u}_1\|_{H_0^1\times L^2}\le E_0,\\ &\|S_k(D)\underline{u}_0\|_{H_0^1\times L^2}\le r, && \|S_k(D)\underline{u}_1\|_{H_0^1\times L^2}\le r \end{aligned} \tag{1.45}$$

there exists a function $g\in L^1(0,T;L^2(\Omega))$ *which exactly controls the semilinear wave equation at time* T*; namely, the unique solution of the system*

$$\begin{cases} \Box u+f(u)=\chi(t,x)g \quad \text{in }]0,+\infty[\times\Omega, \qquad u_{|\partial\Omega}=0,\\ \qquad (u|_{t=0},\partial_t u|_{t=0})=\underline{u}_0\in H_0^1(\Omega)\times L^2(\Omega)\end{cases} \tag{1.46}$$

satisfies $\underline{u}_1=(u|_{t=T},\partial_t u|_{t=T})$*.*

Remark 1.9. The estimate on the values of r and k are deduced from formula (3.22). In particular, one obtains a condition of the form

$$2^{-k\mu}+r\le A(1+E_0)^{-B} \tag{1.47}$$

with $\mu=\frac{5-p}{4p}$ and where the constants $A,B>0$ depend only on Ω and T.

We also have the following corollary.

COROLLARY 1.10. *Assume that* (GCC) *holds true. If* $(\underline{u}_0^n)$ *and* $(\underline{u}_1^n)$ *are two sequences of initial and final data weakly converging to* 0 *in* $H_0^1(\Omega)\times L^2(\Omega)$*, then there exists a sequence of functions* $g_n\in L^1(0,T;L^2(\Omega))$ *which exactly controls the semilinear wave equation at time* T*. More precisely, for* $n\ge n_0$*,* n_0 *large enough, the unique solution of the system*

$$\begin{cases} \Box u_n+f(u_n)=\chi(t,x)g_n \quad \text{in }]0,+\infty[\times\Omega, \qquad u_{n|\partial\Omega}=0,\\ \qquad (u_n|_{t=0},\partial_t u_n|_{t=0})=\underline{u}_0^n\end{cases} \tag{1.48}$$

satisfies $\underline{u}_1^n=(u_n|_{t=T},\partial_t u_n|_{t=T})$*.*

Indeed, there exists E_0 such that $\|\underline{u}_0^n\|_{H_0^1\times L^2}\le E_0$ and $\|\underline{u}_1^n\|_{H_0^1\times L^2}\le E_0$. For every $k\in\mathbb{N}$, it is clear that the dyadic sums $S_k(D)\underline{u}_0^n$ and $S_k(D)\underline{u}_1^n$ strongly converge to 0. Therefore, with the integer k and the real r provided by Theorem 1.8, condition (1.45) is fulfilled for $n \geq n_0$, for n_0 large enough. In particular, the above corollary means that any highly oscillating perturbation of the zero state can be exactly controlled to zero in the time T given by (GCC).

Remarks.

1. Besides (GCC), we assume that the low frequencies of the initial data $\underline{u}_0$ and those of the target $\underline{u}_1$ are small enough in the energy space $H_0^1 \times L^2$. We have to notice that it is an open question to know if the result of Theorem 1.8 holds true without the condition (1.45). The important fact here is that this control is achieved in uniform time, namely, the time T of (GCC). In other words, the control time is the same as the one spent for the control of the linear equation.
2. Concerning the control problem for semilinear wave equations, numerous works are available in the literature (see [Zua90], [Zua93]). Most of them assume the initial data to be small enough in the energy space, and the proofs are then built on a fixed point process. The control time is the one of the linear equation, but, clearly, the general setting is nothing more than a perturbation of the linear case. Another approach is to first dampen the equation in order to decrease the solution energy and then to come back to the previous method. But, in this case, the control time is no longer uniform (see [DLZ03]). Finally, what Theorem 1.8 states is that the semilinear nature of the equation is governed by separation between low and high frequencies. For instance, using again the example of Remark 1.5, for n large enough, (e_n, e_n) can be driven to (e_{n+1}, e_{n+1}) in time T.

2. Linear waves.

2.1. A proof of Theorem 1.3. The proof is achieved in several steps.

Step 1: Q_+ is an isomorphism of L^2. Since (GCC) holds true, we know (see, for example, [BLR92]) that there exists C such that for any solution u of the wave equation $\Box u = 0$ the following observability inequality holds true:

$$\|u(0,.)\|_{L^2}^2 + \|\partial_t u(0,.)\|_{H^{-1}}^2 \leq C \int_0^T \|\chi u(t,x)\|_{L^2}^2 \, dt. \tag{2.1}$$

Let $H = L^2$ and $B = \chi$. Then Q_+ is exactly the operator M_T associated in section 1.1 with the evolution problem $(\partial_t - i\lambda)f = Bg$. Thus Q_+ is an isomorphism of H iff the observability inequality (1.10) holds true. With the notation of (1.10), $u = e^{-is\lambda}h$ is an L^2 solution of the wave equation, and thus (1.10) follows from (2.1).

Step 2: By Lemma 4.2, we know that Q_+ is bounded on $H^s(\Omega, \Delta)$ for all $s \geq 0$. Next we define the operator

$$A_s = \lambda^s Q_+ \lambda^{-s} - Q_+ = \lambda^s [Q_+, \lambda^{-s}]. \tag{2.2}$$

We claim that, for all $s \geq 0$, A_s maps $H^0(\Omega, \Delta) = L^2$ into $H^{1/2}(\Omega, \Delta)$ and therefore is compact on L^2. Indeed, by Lemma 4.3, we know that this holds true for $s \in [0, 2[$. For $s \geq 2$, set $s = 2N + \sigma$ with $\sigma \in [0, 2[$. By integration by parts, one gets

$$\begin{aligned} A_s = \lambda^s Q_+ \lambda^{-s} - Q_+ &= \int_0^T (-\partial_t^2)^N (e^{it\lambda}) \lambda^\sigma \chi^2 \lambda^{-\sigma-2N} e^{-it\lambda} dt - Q_+ \\ &= (-1)^N \sum_{j=0}^{2N-1} \int_0^T e^{it\lambda} P_{j,N} e^{-it\lambda} dt + \lambda^\sigma Q_+ \lambda^{-\sigma} - Q_+ \end{aligned} \tag{2.3}$$

with $P_{j,N} = C^j_{2N}\lambda^\sigma(\partial_t^{2N-j}\chi^2)\lambda^{-2N-\sigma}(-i\lambda)^j$. Then the result follows since, by (4.20), $P_{j,N}$ is bounded from L^2 into $H^a(\Omega,\Delta)$ for any $a < \min(1, 5/2-\sigma)$.

Step 3: Q_+ is an isomorphism on $H^s(\Omega,\Delta)$. Since we know that Q_+ is an isomorphism on L^2, it remains to show the following regularity result for $s \geq 0$:

$$u \in L^2 \text{ and } Q_+u \in H^s(\Omega,\Delta) \Rightarrow u \in H^s(\Omega,\Delta). \tag{2.4}$$

Set $F_s = \{u \in L^2 \text{ and } Q_+u \in H^s(\Omega,\Delta)\}$. Then F_s is a Hilbert space, one has $H^s(\Omega,\Delta) \subset F_s$, and (2.4) is equivalent to $H^s(\Omega,\Delta) = F_s$. Since A_s is a compact operator on L^2, there exists C such that for all $u \in H^s(\Omega,\Delta)$ one has

$$\|u\|_{H^s(\Omega,\Delta)} \leq C(\|Q_+u\|_{H^s(\Omega,\Delta)} + \|u\|_{L^2}). \tag{2.5}$$

In fact, if (2.5) does not hold true, then there exists a sequence $v_n \in L^2$, $\|v_n\| = 1$ such that $\|\lambda^s Q_+ \lambda^{-s} v_n\|_{L^2} + \|\lambda^{-s}v_n\|_{L^2} \to 0$. Then v_n converges weakly to 0, and, since A_s is compact, $\|Q_+v_n\|_{L^2} \to 0$. This contradicts $\|v_n\| = 1$ since Q_+ is an isomorphism on L^2. Thus $H^s(\Omega,\Delta)$ is a closed subspace of F_s. Set $J_\varepsilon = (1+\varepsilon\lambda)^{-1}$. Then $[Q_+, J_\varepsilon] = \varepsilon J_\varepsilon[\lambda, Q_+]J_\varepsilon$ and, by Lemma 4.3, the operator $[\lambda, Q_+]$ is bounded on L^2. For $s < 1$ and $u \in F_s$, one has $J_\varepsilon u \in H^s(\Omega,\Delta)$ and we get

$$\|Q_+J_\varepsilon u - Q_+u\|_{H^s(\Omega,\Delta)} \leq \|(J_\varepsilon - 1)\lambda^s Q_+u\|_{L^2} + \|\lambda^s[Q_+, J_\varepsilon]u\|_{L^2} \to 0 \quad (\varepsilon \to 0).$$

Thus $H^s(\Omega,\Delta)$ is dense in F_s for $s < 1$, and (2.4) holds true for $s < 1$.

To conclude, we proceed by induction on s. Assume that (2.4) holds true on $[0,\sigma]$. Let $0 < s < 1/2$. Then $Q_+u \in H^{s+\sigma}(\Omega,\Delta)$ implies $v = \lambda^\sigma u \in L^2$. Then $\lambda^\sigma Q_+\lambda^{-\sigma}v \in H^s(\Omega,\Delta)$ gives $Q_+v + A_\sigma v \in H^s(\Omega,\Delta)$, and one has $A_\sigma v \in H^{1/2}(\Omega,\Delta)$ by Step 2. Thus we get $Q_+v \in H^s(\Omega,\Delta)$. Since $s < 1$, this implies $v \in H^s(\Omega,\Delta)$, which gives $u \in H^{s+\sigma}(\Omega,\Delta)$. Thus (2.4) holds true for any $s \geq 0$.

To conclude the proof, we now have to verify that the operators $\mathcal{T}$ and $\mathcal{T}^*$ are smoothing. We proceed by a complex deformation. Take N large and set $\theta(t+is,x) = \sum_{0\leq j\leq N} \frac{(is)^j}{j!}\partial_t^j(\chi^2(t,x))$. Let c be a complex path connecting 0 to T in $\mathrm{Im}(z) > 0$, with small and no zero angles with the real axis at 0 and T, and let U be the open set in $\mathbb{C}$ with boundary $[0,T] \cup c$. Let $\alpha(z,x) = e^{iz\lambda}\theta(z,x)e^{iz\lambda}dz$. By the Stokes formula, one has

$$\mathcal{T} = \int_U d\alpha + \int_c \alpha. \tag{2.6}$$

Let $a > 0$ be given. For $\mathrm{Im}(z) > 0$, $e^{iz\lambda}$ maps L^2 into $H^a(\Omega,\Delta)$ with norm $\leq C\,\mathrm{Im}(z)^{-a}$. Since the smooth function $\chi^2(t,x)$ is flat at $t = 0$ and $t = T$, for any $M > 0$, there exists a constant C_M such that one has $|\theta(z)| \leq C_M dist(z, \{0,T\})^M$ for any $z \in U$. Since $\mathrm{Im}(z) \geq C dist(z,\{0,T\})$ on the complex path c, we get that $\int_c \alpha$ maps L^2 into $H^a(\Omega,\Delta)$. On the other hand, $d\alpha = e^{iz\lambda}\overline{\partial_z}\theta(z,x)e^{iz\lambda}d\overline{z}dz$ and one has $\left|\overline{\partial_z}\theta(z,x)\right| \leq C_N\ |\mathrm{Im}(z)|^N$ for $z \in U$, and therefore $\int_U d\alpha$ maps L^2 into $H^a(\Omega,\Delta)$ for $a \leq N$. Thus $\mathcal{T}$ is smoothing, and the same argument using a complex deformation in $\mathrm{Im}(z) < 0$ shows that $\mathcal{T}^*$ is smoothing. Finally, (1.35) is now an obvious consequence of (1.31), (1.32), and $\Lambda = M_T^{-1}$. The proof of Theorem 1.3 is complete.

2.2. A proof of Theorem 1.4. The proof of Theorem 1.4 is an easy by-product of the proof of Theorem 1.3. One has, with $h = 2^{-k}$, $\psi_k(D) = \psi(h|\Delta|^{1/2}) = A_h$, with

$\psi \in C_0^\infty$ and supported in $[1/4, 1]$. Thus there exists C such that one has for all $h \in]0, 1]$

$$\begin{aligned} \forall u \in L^2 \quad \|A_h u\|_{H^{-1}(\Omega,\Delta)} &\leq Ch\|u\|_{L^2}, \\ \forall u \in H^1(\Omega,\Delta) \quad \|A_h u\|_{L^2} &\leq Ch\|u\|_{H^1(\Omega,\Delta)} \end{aligned} \tag{2.7}$$

and by (4.37)

$$\|[A_h, \chi_0^2]u\|_{L^2} \leq Ch\|u\|_{L^2}. \tag{2.8}$$

The first line of (1.38) is equivalent to the estimate

$$\forall h \in]0, 1] \quad \|[A_h, \Lambda]\|_{L^2} \leq Ch. \tag{2.9}$$

Since $[A_h, \Lambda] = \Lambda[M_T, A_h]\Lambda$ and since Λ is an isomorphism on L^2, we have to prove $\|[M_T, A_h]\|_{L^2} \leq Ch$. Let B be an operator bounded from L^2 into $H^1(\Omega, \Delta)$ and from $H^{-1}(\Omega, \Delta)$ into L^2. Then by (2.7) one has $\|BA_h\|_{L^2} \leq Ch$ and $\|A_hB\|_{L^2} \leq Ch$. Observe that, thanks to (1.31) and (1.32), M_T is given by formula

$$M_T = \frac{1}{2}\begin{pmatrix} Q_+ & -\mathcal{T} \\ -\mathcal{T}^* & Q_- \end{pmatrix}. \tag{2.10}$$

The operators $\mathcal{T}$ and $\mathcal{T}^*$ are of type B by Theorem 1.3. Thus it remains to show that

$$\forall h \in]0, 1] \quad \|[A_h, Q_\pm]\|_{L^2} \leq Ch, \tag{2.11}$$

which is obvious by definition (1.33) of $Q_\pm$, (2.8), and $[\lambda, A_h] = 0$. We get the second line of (1.38) from the first line since by (1.37) one has $S_k(D) = 1 - \sum_{j>k} \psi_j(D)$. The proof of Theorem 1.4 is complete.

3. Semilinear waves. This section is devoted to the proof of Theorem 1.8. Without loss of generality, in what follows we may assume that $4 \leq p < 5$ since the hypothesis (1.40) on the nonlinearity f is monotone in p. First, we decompose the semilinear term

$$f(u) = f'(0)u + \theta(u),$$

where θ is of class C^3 and satisfies for some $c > 0$

$$|\theta(s)| \leq c(|s|^2 + |s|^p). \tag{3.1}$$

In the proof of Theorem 1.8, we will use the fact that the problem is subcritical and therefore close to a linear control problem thanks to the hypothesis (1.45). Therefore, we will choose the control function g in (1.46) in the form $g = \chi(t, x)(g_1 + g_2)$, where both g_1 and g_2 are solutions of the linear equation

$$\begin{aligned} \Box g_1 + f'(0)g_1 = 0, \quad g_{1|\partial\Omega} = 0 \text{ and } (g_1|_{t=0}, \partial_t g_1|_{t=0}) \in L^2 \times H^{-1}, \\ \Box g_2 + f'(0)g_2 = 0, \quad g_{2|\partial\Omega} = 0 \text{ and } (g_2|_{t=0}, \partial_t g_2|_{t=0}) \in L^2 \times H^{-1}. \end{aligned} \tag{3.2}$$

We choose for g_1 the linear HUM control associated with the operator $\Box + f'(0)$ and with the initial data $\underline{u}_0$ and the final data $\underline{u}_1$, so our unknown will be $\underline{g}_2 = (g_2|_{t=0}, \partial_t g_2|_{t=0})$.

Now consider the system

$$
\tag{3.3}
\left\{
\begin{array}{c}
\Box u + f(u) = \chi^2(t,x)g_1 + \chi^2(t,x)g_2\,, \quad u_{|\partial\Omega} = 0,\\
(u|_{t=0}, \partial_t u|_{t=0}) = \underline{u}_0 \in H_0^1 \times L^2.
\end{array}
\right.
$$

For every $g_2 \in L^1(0,T;L^2(\Omega))$, system (3.3) admits a unique solution u in the space $C^0(0,T;H_0^1(\Omega))$. Moreover, u satisfies the Strichartz estimates (1.42), (1.43). We split u into $u = v + w$ with

$$
\tag{3.4}
\Box v + f'(0)v = \chi^2(t,x)g_1, \quad v_{|\partial\Omega} = 0 \text{ and } (v|_{t=0}, \partial_t v|_{t=0}) = \underline{u}_0.
$$

Clearly, by our choice of g_1, we have $(v|_{t=T}, \partial_t v|_{t=T}) = \underline{u}_1$, and thus w must satisfy $(w|_{t=T}, \partial_t w|_{t=T}) = 0$ and

$$
\tag{3.5}
\Box w + f'(0)w = -\theta(u) + \chi^2(t,x)g_2\,, \quad w_{|\partial\Omega} = 0 \text{ and } (w|_{t=0}, \partial_t w|_{t=0}) = 0.
$$

Let h be the solution of

$$
\tag{3.6}
\Box h + f'(0)h = \theta(u), \quad h_{|\partial\Omega} = 0 \text{ and } (h|_{t=T}, \partial_t h|_{t=T}) = 0.
$$

We denote by $\mathcal{N}$ the map from $L^2 \times H^{-1}$ into $H_0^1 \times L^2$ defined by

$$
\tag{3.7}
\mathcal{N}(\underline{g}_2) = (h|_{t=0}, \partial_t h|_{t=0}),
$$

where h is the solution of (3.6) and u is the solution of (3.2). The function $k = w + h$ solves the system

$$
\tag{3.8}
\Box k + f'(0)k = \chi^2(t,x)g_2\ , \quad k_{|\partial\Omega} = 0 \text{ and } (k|_{t=0}, \partial_t k|_{t=0}) = (h|_{t=0}, \partial_t h|_{t=0}).
$$

If there exists $\underline{g}_2 \in L^2 \times H^{-1}$ such that $\mathcal{N}(\underline{g}_2) = (h|_{t=0}, \partial_t h|_{t=0})(k|_{t=0}, \partial_t k|_{t=0}) = \Lambda^{-1}(\underline{g}_2)$, where Λ is the linear HUM control associated with the operator $\Box + f'(0)$ (except that we have exchanged the role of 0 and T), then by formulae (3.3), (3.4), (3.5), (3.6), and (3.8) the control $g = \chi(t,x)(g_1 + g_2)$ in (1.46) will drive the initial state $\underline{u}_0$ to the final state $\underline{u}_1$.

Hence, we are looking for a fixed point for the operator

$$
L = \Lambda\mathcal{N}.
$$

Actually, we will prove that, under the conditions of Theorem 1.8, L reproduces a small ball B_ρ centered at the origin of $L^2 \times H^{-1}$ and is contracting on B_ρ. In fact, taking advantage of the regularity of the composition $f(u)$, we easily see that L is compact; thus, we could argue as in the proof of Theorem 3 of [DLZ03] and use a Schauder fixed point result. But keeping in mind the fact that the present result could have numerical applications, we will instead run a classical Picard fixed point theorem.

We denote by $C(E_0)$ constants which depend only on E_0 and by c various constants independent of $\underline{u}_0, \underline{u}_1$. To begin, take $\underline{g}_2 \in B_\rho$ ($\rho < 1$), $k \in \mathbb{N}$, and let $S_k(D)$ be the spectral localization defined in (1.37). One has

$$
\left\| L\underline{g}_2 \right\|_{L^2 \times H^{-1}} \le c \left\| \mathcal{N}\underline{g}_2 \right\|_{H_0^1 \times L^2} = c \left\| (h|_{t=0}, \partial_t h|_{t=0}) \right\|_{H_0^1 \times L^2}.
$$

Thanks to the hyperbolic energy estimate applied to (3.6), one has

$$
\left\| (h|_{t=0}, \partial_t h|_{t=0}) \right\|_{H_0^1 \times L^2} \le c \int_0^T \left(\int_\Omega |\theta(u)|^2\, dx \right)^{1/2} dt.
$$

Or, equivalently,

$$\|(h|_{t=0}, \partial_t h|_{t=0})\|_{H_0^1 \times L^2} \leq c \int_0^T \|u(t)\|_{L^4}^2 \, dt + c \int_0^T \|u(t)\|_{L^{2p}}^p \, dt. \tag{3.9}$$

We split u into

$$u = u_L + u_H,$$

where

$$u_L = S_k(D)u \qquad \text{and} \qquad u_H = (1 - S_k(D))u$$

are, respectively, the low and high frequency parts of u. This gives

$$\left\{ \begin{aligned} \int_0^T \left(\int_\Omega |\theta(u)|^2 \, dx \right)^{1/2} dt \leq c \int_0^T \|u_H(t)\|_{L^4}^2 \, dt + c \int_0^T \|u_H(t)\|_{L^{2p}}^p \, dt \\ + c \int_0^T \|u_L(t)\|_{L^4}^2 \, dt + c \int_0^T \|u_L(t)\|_{L^{2p}}^p \, dt. \end{aligned} \right. \tag{3.10}$$

We will successively estimate these integrals.

First, applying $(1 - S_k(D))$ to (3.3), we obtain

$$\Box u_H = (1 - S_k(D)) \left[-f(u) + \chi^2(g_1 + g_2) \right] \in L^1(0, T; L^2(\Omega)).$$

Consequently, $\|u_H\|_{L^5(0,T;W_0^{\frac{3}{10},5})}$ and $\|u_H\|_{L^\infty(0,T;H_0^1)}$ are bounded by $C(E_0)$ as well as all of the Strichartz norms $\|u_H\|_{L^a(0,T;L^{3r})}$, $a \geq 5$. (Notice that the same fact holds for u_L.) In particular, one has, for all $t \in [0, T]$, $\|u_H\|_{L^2} \leq C(E_0) 2^{-k}$ and $\|u_H\|_{L^5(0,T;L^{10})} \leq C(E_0)$. For $q \in [1, 5[$, one has $\|u_H(t)\|_{L^{2q}} \leq \|u_H(t)\|_{L^2}^\theta \|u_H(t)\|_{L^{10}}^{1-\theta}$ with $\theta = \frac{5-q}{4q}$, and thus we get

$$\|u_H(t)\|_{L^{2q}}^q \leq 2^{-k(\frac{5-q}{4})} \|u_H(t)\|_{L^{10}}^{\frac{5}{4}(q-1)},$$

and this yields

$$\int_0^T \|u_H(t)\|_{L^{2q}}^q \, dt \leq C(E_0) 2^{-k(\frac{5-q}{4})}. \tag{3.11}$$

Thus we obtain with $\mu = \frac{5-p}{4p}$

$$\int_0^T \|u_H(t)\|_{L^4}^2 \, dt + \int_0^T \|u_H(t)\|_{L^{2p}}^p \, dt \leq C(E_0) 2^{-kp\mu}. \tag{3.12}$$

It remains to estimate the last two integrals of (3.10), which are the contribution of the low frequencies of u. For this purpose, we examine the system satisfied by u_L.

Applying $S_k(D)$ to (3.3), we obtain

$$\left\{ \begin{aligned} \Box u_L + f(u_L) = [\theta(u_L) - S_k(D)\theta(u)] + S_k(D)[\chi^2 g_1 + \chi^2 g_2], \\ (u_L(0), \partial_t u_L(0)) = S_k(D)\underline{u}_0 \in H_0^1 \times L^2. \end{aligned} \right. \tag{3.13}$$

Now, writing $\theta(u) - \theta(u_L) = u_H \theta'(u_L + \alpha u_H)$, $0 < \alpha < 1$, and taking into account assumption (1.40), we get

$$|\theta(u) - \theta(u_L)| \leq c\,|u_H|\,(|u_L| + |u_H| + |u_L|^{p-1} + |u_H|^{p-1}).$$

We just estimate the term $|u_H|\,|u_L|^{p-1}$, the others being simpler. By the Holder inequality, we have

$$\int_\Omega |u_H|^2\,|u_L|^{2(p-1)}\,dx \leq \|u_H\|_{L^{2p}}^2\,\|u_L\|_{L^{2p}}^{2(p-1)}$$

and we deduce that

$$\int_0^T \left(\int_\Omega |u_H|^2\,|u_L|^{2(p-1)}\,dx\right)^{1/2} dt \leq \left(\int_0^T \|u_H(t)\|_{L^{2p}}^5\,dt\right)^{1/5} \cdot \left(\int_0^T \|u_L(t)\|_{L^{2p}}^{\frac{5}{4}(p-1)}\,dt\right)^{4/5}.$$

By Strichartz estimates, the second integral of the right-hand side is bounded by $C(E_0)$ since $\frac{5}{4}(p-1) < p$. Moreover, arguing as in (3.11), we get

$$\|u_H\|_{L^5(0,T;L^{2p})} \leq C(E_0)2^{-k\mu}.$$

This yields

$$\theta(u_L) = \theta(u) + 0(2^{-k\mu}).$$

Here we denote by $0(2^{-k\mu})$ any function g such that

$$\int_0^T \left(\int_\Omega |g|^2\,dx\right)^{1/2} dt \leq C(E_0)2^{-k\mu}.$$

Consequently,

$$\theta(u_L) - S_k(D)\theta(u) = (1 - S_k(D))\theta(u) + 0(2^{-k\mu}).$$

Extending u by zero outside Ω and using the regularity of the composition (Theorem 4.7), one gets $\theta(u) \in L^1(0,T;\,H^{\frac{3}{10}(5-p)})$. Since $\theta(u)|_{\partial\Omega} = 0$, $\frac{3}{10}(5-p) > \mu$, and $\mu < 1$, we get $\theta(u) \in L^1(0,T;H^\mu(\triangle,\Omega))$. Thus we conclude that the bracket in the right-hand side of (3.13) satisfies

$$\theta(u_L) - S_k(D)\theta(u) = 0(2^{-k\mu}). \tag{3.14}$$

On the other hand, inequality (1.38) of Theorem 1.4 and (4.37) give for every $t \in [0,T]$

$$\begin{aligned}\left\|S_k(D)(\chi^2(t,x)g_1(t))\right\|_{L^2} &\leq \left\|\chi^2(t,x)S_k(D)g_1(t)\right\|_{L^2} + \left\|[S_k(D),\chi^2(t,x)]g_1(t)\right\|_{L^2}\\ &\leq c\left\|S_k(D)\underline{g}_1\right\|_{L^2\times H^{-1}} + c2^{-k}\left\|\underline{g}_1\right\|_{L^2\times H^{-1}}\\ &\leq c\left(\|S_k(D)\underline{u}_0\|_{H^1\times L^2} + \|S_k(D)\underline{u}_1\|_{H^1\times L^2}\right)\\ &\quad + c2^{-k}\left(\|\underline{u}_0\|_{H^1\times L^2} + \|\underline{u}_1\|_{H^1\times L^2}\right).\end{aligned}$$

Hence, by hypothesis (1.45) of Theorem 1.8, we get

$$\left\|S_k(D)(\chi^2(t,x)g_1)\right\|_{L^1(0,T;L^2)} \leq cr + cE_0 2^{-k}. \tag{3.15}$$

And obviously, since g_2 satisfies (3.2),

$$\left\|S_k(D)(\chi^2(t,x)g_2)\right\|_{L^1(0,T;L^2)} \leq \left\|\chi^2(t,x)g_2\right\|_{L^1(0,T;L^2)} \leq c\rho. \tag{3.16}$$

Summarizing (3.14), (3.15), (3.16), using hypothesis (1.45) and $\mu < 1$, we obtain the hyperbolic energy estimate

$$\sup_{0\leq t\leq T}(Eu_L(t))^{1/2} \leq (Eu_L(0))^{1/2} + C(E_0)(r + 2^{-k\mu} + \rho) \leq C(E_0)(r + 2^{-k\mu} + \rho),$$

and this gives, in particular,

$$\sup_{0\leq t\leq T}\|u_L(t)\|_{H_0^1} \leq C(E_0)(r + 2^{-k\mu} + \rho). \tag{3.17}$$

On the other hand, thanks to assumption (1.40), we have

$$\|f(u_L)\|_{L^1(0,T;L^2)} \leq c\,\|u_L\|_{L^1(0,T;L^2)} + c\,\||u_L|^p\|_{L^1(0,T;L^2)}\,.$$

Combined with (3.17), (3.13), and (4.56), this yields the Strichartz inequality

$$\|u_L\|_{L^5(0,T;W^{\frac{3}{10},5})} \leq C(E_0)(r + 2^{-k\mu} + \rho) + C(E_0)\,\|u_L\|^p_{L^5(0,T;W^{\frac{3}{10},5})}\,. \tag{3.18}$$

Of course, in this estimate we can replace the time T by any value $t \in [0,T]$. Thus, applying the boot-strap lemma, we infer that

$$\|u_L\|_{L^5(0,T;W^{\frac{3}{10},5})} \leq 2C(E_0)(r + 2^{-k\mu} + \rho) \tag{3.19}$$

if

$$(r + 2^{-k\mu} + \rho) < (2C(E_0))^{-\frac{p}{p-1}}. \tag{3.20}$$

Then we deduce that

$$\begin{aligned}&\int_0^T \|u_L(t)\|^2_{L^4}\,dt + \int_0^T \|u_L(t)\|^p_{L^{2p}}\,dt \\ &\leq C(E_0)\,\|u_L\|^2_{L^5(0,T;W^{\frac{3}{10},5})} \leq C'(E_0)(r + 2^{-k\mu} + \rho)^2.\end{aligned} \tag{3.21}$$

Finally, plugging (3.12) and (3.21) into (3.10), we conclude that

$$\left\|L\underline{g}_2\right\|_{L^2\times H^{-1}} \leq c\,\|(h|_{t=0}, \partial_t h|_{t=0})\|_{H^1\times L^2} \leq C'(E_0)(r + 2^{-k\mu} + \rho)^2 + C(E_0)2^{-kp\mu}.$$

Thus there exists a constant $C(E_0)$ such that

$$\left\|L\underline{g}_2\right\|_{L^2\times H^{-1}} \leq C(E_0)(r + 2^{-k\mu} + \rho)^2. \tag{3.22}$$

Then one can easily check that there exists a constant $\rho(E_0) > 0$ such that, for any $\rho \leq \rho(E_0)$ and any r, k such that $r + 2^{-k\mu} \leq \rho$, condition (3.20) is satisfied and

$$\left\|L\underline{g}_2\right\|_{L^2\times H^{-1}} \leq \rho.$$

Now it remains to prove that, if ρ is small enough, L is contracting on B_ρ. For that, we examine the systems

$$\Box u_j + f(u_j) = \chi^2(t,x)g_1 + \chi^2(t,x)g_2^j,$$

$$u_{j|\partial\Omega} = 0 \text{ and } (u_j|_{t=0}, \partial_t u_j|_{t=0}) = \underline{u}_0 \in H_0^1 \times L^2,$$

$$\Box h_j + f'(0)h_j = \theta(u_j), \quad h_{j|\partial\Omega} = 0 \text{ and } (h_j|_{t=0}, \partial_t h_j|_{t=0}) = 0$$

for $j = 1, 2$, where the data $\underline{g}_2^1$, $\underline{g}_2^2$ are in the ball B_ρ of $L^2 \times H^{-1}$. We have

$$\tag{3.23} \left\{ \begin{array}{c} \Box(h_1 - h_2) + f'(0)(h_1 - h_2) = \theta(u_1) - \theta(u_2), \\ ((h_1 - h_2)(T), \partial_t(h_1 - h_2)(T)) = (0,0) \end{array} \right.$$

and

$$\tag{3.24} \left\{ \begin{array}{c} \Box(u_1 - u_2) + f(u_1) - f(u_2) = \chi^2(t,x)(g_2^1 - g_2^2), \\ ((u_1 - u_2)(0), \partial_t(u_1 - u_2)(0)) = (0,0). \end{array} \right.$$

We estimate

$$\tag{3.25} \begin{aligned} &\left\| L\underline{g}_2^1 - L\underline{g}_2^2 \right\|_{L^2 \times H^{-1}} \leq c \left\| (h_1|_{t=0}, \partial_t h_1|_{t=0}) - (h_2|_{t=0}, \partial_t h_2|_{t=0}) \right\|_{H^1 \times L^2} \\ &\leq c \int_0^T \|\theta(u_1) - \theta(u_2)\|_{L^2} \, dt. \end{aligned}$$

Writing $\theta(u_1) - \theta(u_2) = (u_1 - u_2)\ \theta'(u_1 + \alpha u_2)$, $0 < \alpha < 1$, we get

$$\tag{3.26} \begin{aligned} &\left\| L\underline{g}_2^1 - L\underline{g}_2^2 \right\|_{L^2 \times H^{-1}} \\ &\qquad \leq c \int_0^T \left(\int_\Omega |u_1 - u_2|^2 \left(|u_1|^2 + |u_2|^2 + |u_1|^{2(p-1)} + |u_2|^{2(p-1)} \right) dx \right)^{1/2} dt. \end{aligned}$$

As before, we just examine the integral corresponding to the term $|u_1 - u_2|^2 \, |u_1|^{2(p-1)}$. We have

$$\begin{aligned} \int_0^T \left(\int_M |u_1 - u_2|^2 \, |u_1|^{2(p-1)} \, dx \right)^{1/2} dt &\leq \|u_1 - u_2\|_{L^5(0,T;L^{2p})} \cdot \|u_1\|^{(p-1)}_{L^{\frac{5}{4}(p-1)}(0,T;L^{2p})} \\ &\leq \|u_1 - u_2\|_{L^5(0,T;W^{\frac{3}{10},5})} \cdot \|u_1\|^{(p-1)}_{L^{\frac{5}{4}(p-1)}(0,T;L^{2p})} . \end{aligned}$$

Moreover,

$$\tag{3.27} \|u_1\|^{(p-1)}_{L^{\frac{5}{4}(p-1)}(0,T;L^{2p})} \leq C(T) \, \|u_1\|^{p-1}_{L^p(0,T;L^{2p})} .$$

We decompose $u_1 = u_{1L} + u_{1H}$, and, following the arguments of the beginning of the proof, we obtain an estimate similar to (3.12)

$$\tag{3.28} \|u_{1H}\|_{L^p(0,T;L^{2p})} \leq C(E_0) 2^{-k\mu} \leq C(E_0)\rho$$

and a second one similar to (3.19)

$$\|u_{1L}\|_{L^p(0,T;L^{2p})} \leq C(T) \left\|u_{1L}\right\|_{L^5(0,T;W^{\frac{3}{10},5})} \leq C(T,E_0)\rho. \tag{3.29}$$

Therefore, we get from (3.26)

$$\left\|L\underline{g}_2^1 - L\underline{g}_2^2\right\|_{L^2\times H^{-1}} \leq C(T,E_0)\rho \left\|u_1 - u_2\right\|_{L^5(0,T;W^{\frac{3}{10},5})}. \tag{3.30}$$

On the other hand, the hyperbolic energy estimate applied to system (3.24) gives

$$\begin{aligned}\sup_{0\leq t\leq T} \|(u_1-u_2)(t)\|_{H^1} &\leq c\int_0^T \|\theta(u_1)-\theta(u_2)\|_{L^2}\,dt + c\int_0^T \left\|\chi^2(g_2^1-g_2^2)\right\|_{L^2} dt \\ &\leq C(T,E_0)\rho\|u_1-u_2\|_{L^5(0,T;W^{\frac{3}{10},5})} + c\int_0^T \left\|\chi^2(g_2^1-g_2^2)\right\|_{L^2} dt.\end{aligned}$$

Combining then with the Strichartz inequality applied to (3.24), we obtain

$$\begin{aligned}\|u_1-u_2\|_{L^5(0,T;W^{\frac{3}{10},5})} &\leq c\int_0^T \|f(u_1)-f(u_2)\|_{L^2}\,dt + c\int_0^T \left\|\chi^2(g_2^1-g_2^2)\right\|_{L^2} dt \\ &\leq C(T,E_0)\rho\left\|u_1-u_2\right\|_{L^5(0,T;W^{\frac{3}{10},5})} + c\int_0^T \left\|\chi^2(g_2^1-g_2^2)\right\|_{L^2} dt,\end{aligned}$$

which, in turn, gives

$$\|u_1-u_2\|_{L^5(0,T;W^{\frac{3}{10},5})} \leq c\int_0^T \left\|\chi^2(g_2^1-g_2^2)\right\|_{L^2} dt$$

for ρ small enough. Plugging this estimate into (3.30), we finally get

$$\begin{aligned}\left\|L\underline{g}_2^1 - L\underline{g}_2^2\right\|_{L^2\times H^{-1}} &\leq C(T,E_0)\rho\int_0^T \left\|\chi^2(g_2^1-g_2^2)(t)\right\|_{L^2} dt \\ &\leq C(T,E_0)\rho\left\|\underline{g}_2^1-\underline{g}_2^2\right\|_{L^2\times H^{-1}}.\end{aligned}$$

This shows that the operator L is contracting on B_ρ for a suitable choice of ρ and concludes the proof of Theorem 1.8.

4. Appendix.

4.1. The case of a compact manifold. In this section, we study the case of internal control on a compact manifold M without boundary. The situation here is much simpler, and we will get that the optimal control map Λ is a zero order elliptic pseudodifferential operator in Theorem 4.1.

Let us recall that the problem of controllability is stated in the following way. Let T be a positive time and $\chi(t,x)$ be as in (1.19). For a given $(u_0,u_1) \in H = H^1(M)\times L^2(M)$, the problem is to find a source $v(t,x) \in L^2(0,T;L^2(M))$ such that the solution of the system

$$\left\{\begin{array}{l}\Box u = (\partial_t^2-\Delta)u = \chi v \quad \text{in }]0,+\infty[\times M, \\ (u(0),\partial_t u(0)) = (0,0)\end{array}\right. \tag{4.1}$$

reaches the state $(u(T), D_t u(T)) = (u_0, u_1)$ at time T. As before, the HUM method consists in taking the control function v in (4.1) in the form $v = \chi w$, where w is a solution of the dual problem

$$\text{(4.2)} \qquad \begin{cases} \Box w = 0 \quad \text{in }]0, +\infty[\times M, \\ (w(0), \partial_t w(0)) = (w_0, w_1) \in L^2 \times H^{-1}. \end{cases}$$

In this geometry, the situation is slightly different from the one we have previously studied in the case $M = \Omega$, since the usual scalar product on $H^1(M)$ is $\int_M (\nabla u \overline{\nabla v} + u\overline{v}) dx$, and so the operator A given by (1.23) and $D(A) = \{\underline{u} \in H,\ A(\underline{u}) \in H\}$ is not self-adjoint. Moreover, $\lambda = \sqrt{-\Delta}$ is no longer an isomorphism since $\lambda(1) = 0$. Thus, we will work with a more convenient and natural scalar product as follows. Let

$$\text{(4.3)} \qquad \begin{gathered} L^2_+ = \left\{ \sum_{j \geq 1} a_j e_j,\ (a_j) \in l^2 \right\} = \left\{ f \in L^2(M), \quad \int_M f = 0 \right\}, \\ \Pi_0(f) = \frac{1}{\sqrt{vol(M)}} \int_M f = (f | e_0), \quad \Pi_+(f) = f - \Pi_0(f) e_0. \end{gathered}$$

Let $\underline{\lambda}$ be the isomorphism of H^1_+ onto L^2_+ given by formula (1.18) restricted to the indices $j \geq 1$. We set

$$\text{(4.4)} \qquad \begin{gathered} H = H^1(M) \times L^2(M) = H_+ \oplus \mathbb{C}^2, \\ H_+ = H^1_+(M) \times L^2_+(M). \end{gathered}$$

Then A is diagonal in the splitting $H_+ \oplus \mathbb{C}^2$, and one has

$$\text{(4.5)} \qquad A = \begin{pmatrix} A_+ & 0 \\ 0 & A_0 \end{pmatrix}, \qquad iA_+ = \begin{pmatrix} 0 & \mathrm{Id} \\ -\underline{\lambda}^2 & 0 \end{pmatrix}, \qquad iA_0 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Then we choose on $H = H^1(M) \times L^2(M)$ the scalar product given by the splitting (4.4), which gives a norm equivalent to the usual one and is such that A_+ is self-adjoint as in the case $M = \Omega$. One has

$$\text{(4.6)} \qquad e^{itA} = \begin{pmatrix} e^{itA_+} & 0 \\ 0 & e^{itA_0} \end{pmatrix}, \qquad e^{itA_0} = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}.$$

In the splitting $H^1(M) = H^1_+(M) \oplus \mathbb{C}$, let $\widetilde{\lambda}$ be the operator

$$\text{(4.7)} \qquad \widetilde{\lambda} = \begin{pmatrix} \underline{\lambda} & 0 \\ 0 & 1 \end{pmatrix}.$$

Then $\widetilde{\lambda}$ is an isomorphism from $H^1(M)$ onto $L^2(M)$, and our scalar product on $H^1(M)$ is given by $(u|v) = \int_M \widetilde{\lambda}(u)\overline{\widetilde{\lambda}(v)}$. Thus, if we set as in (1.25)

$$\text{(4.8)} \qquad B = \begin{pmatrix} 0 & 0 \\ \chi\widetilde{\lambda} & 0 \end{pmatrix},$$

then we still get that the optimal control $g_0(t) = B^* e^{-i(T-t)A^*}\Lambda(f)$ is of the form $g_0 = (\widetilde{\lambda}^{-1}\chi w, 0)$, where w is a solution of (4.2) since, by (4.6), $(e^{itA} - e^{itA^*})f$ belongs to the kernel of $\triangle$ and ∂_t^2. Thus, with this choice of scalar product, we get also in the case where M is compact that the optimal control operator Λ is the inverse of the map M_T given in (1.27). For $\underline{g} = (g_0, g_1) \in H$, set

$$
\begin{aligned}
g_0 &= \underline{\lambda}^{-1}(h_+ + h_-) + c_0 e_0,\\
g_1 &= i(h_+ - h_-) + c_1 e_0.
\end{aligned}
\tag{4.9}
$$

One has $h_\pm \in L^2_+(M)$, $(c_0, c_1) \in \mathbb{C}^2$, and (4.9) gives an identification by an elliptic pseudodifferential operator between the two Hilbert spaces H and $L^2_+(M) \oplus L^2_+(M) \oplus \mathbb{C}^2$.

Set $\mathcal{U}_\pm(t) = e^{\pm it\underline{\lambda}}$. Then the operators $\mathcal{U}_\pm(t)$ are isometries on each Sobolev space $H^s_+(M)$, with inverses $\mathcal{U}_\pm(-t)$. With this identification, one can easily derive the following:

$$
e^{itA} = \left(\mathcal{U}_+(t), \mathcal{U}_-(t), \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}\right).
\tag{4.10}
$$

Moreover, with the notation $\underline{\chi} = \Pi_+ \chi \Pi_+$, the control operator B is given by

$$
B\begin{pmatrix} h_+ \\ h_- \\ c_0 \\ c_1 \end{pmatrix} = \begin{pmatrix} \mathcal{A} & \mathcal{C}' \\ \mathcal{C} & \mathcal{D} \end{pmatrix}\begin{pmatrix} h_+ \\ h_- \\ c_0 \\ c_1 \end{pmatrix},
\tag{4.11}
$$

$$
\begin{aligned}
\mathcal{A} &= \tfrac{1}{2i}\begin{pmatrix} \underline{\chi} & \underline{\chi} \\ -\underline{\chi} & -\underline{\chi} \end{pmatrix}, \qquad \mathcal{C} = \begin{pmatrix} 0 & 0 \\ \Pi_0\chi & \Pi_0\chi \end{pmatrix},\\
\mathcal{C}' &= \begin{pmatrix} a & 0 \\ -a & 0 \end{pmatrix}, \qquad \mathcal{D} = \begin{pmatrix} 0 & 0 \\ b & 0 \end{pmatrix}
\end{aligned}
\tag{4.12}
$$

with $a(t,.) = \frac{1}{2i}\Pi_+(\chi(t,.)e_0) \in L^2_+$, and $b(t,.) = (\chi(t,.)e_0|e_0) \in]0,\infty[$. Of course, $\mathcal{A}$ contains all of the infinite-dimensional part of the control operator B. One has $a \in C^\infty(\mathbb{R}_t, H^s_+(M))$ for all s, and thus $\mathcal{C}'(t,.)$ maps $\mathbb{C}^2$ into $H^\infty_+(M) \oplus H^\infty_+(M)$.

In this splitting, the optimal control map Λ is the inverse of the map M_T given by

$$
M_T = \int_0^T m(s)ds, \quad m(s) = e^{isA}B(T-s)B^*(T-s)e^{-isA^*}.
$$

Let us denote by $\mathcal{R} = \mathcal{R}(t)$ any smooth in t family of operators which maps $H^\sigma_+(M,\Delta)$ into $H^s_+(M,\Delta)$ for all $\sigma, s \in \mathbb{R}$. From (4.11) and (4.12), one has

$$
m(s) = \frac{1}{2}\Pi_+\begin{pmatrix} \mathcal{U}_+(s)(\underline{\chi})^2\mathcal{U}_+(-s) & -\mathcal{U}_+(s)(\underline{\chi})^2\mathcal{U}_-(-s) \\ -\mathcal{U}_-(s)(\underline{\chi})^2\mathcal{U}_+(-s) & \mathcal{U}_-(s)(\underline{\chi})^2\mathcal{U}_-(-s) \end{pmatrix}\Pi_+ + \mathcal{R}.
\tag{4.13}
$$

Let $Q_\pm$ be the operators

$$
Q_\pm = \int_0^T \mathcal{U}_\pm(s)(\underline{\chi})^2\mathcal{U}_\pm(-s)ds,
\tag{4.14}
$$

and let $\mathcal{T}$ be the operator

$$\mathcal{T} = \int_0^T \mathcal{U}_+(s)(\underline{\chi})^2 \mathcal{U}_+(s) ds. \tag{4.15}$$

Let $s \to (\gamma_{(x,\xi)}(s), t - s\tau, \tau)$, $s \in \mathbb{R}$, be the bicharacteristic ray of the wave operator, issued from $(x, \xi, t, \tau = |\xi|_x) \in T^*M \backslash 0$. Let $\alpha(x, \xi)$ be the symbol of order 0 on $T^*M \backslash 0$:

$$\alpha(x,\xi) = \left(\int_0^T \chi^2(T - s, \gamma_{(x,\xi)}(s)) ds \right)^{-1}. \tag{4.16}$$

Obviously, (GCC) guarantees that it is well defined. Moreover, it is homogeneous of degree 0, positive, and elliptic. We will also use the symbol

$$\beta(x,\xi) = \alpha(x, -\xi) = \left(\int_0^T \chi^2(T - s, \gamma_{(x,-\xi)}(s)) ds \right)^{-1}. \tag{4.17}$$

Observe that under (GCC) and thanks to the Egorov theorem the operators $Q_\pm$ defined by (4.14) are elliptic pseudodifferential operators on M, with principal symbol $\sigma(Q_\pm) = (\alpha(x, \pm\xi))^{-1}$.

Theorem 4.1. *Assume that the* (GCC) *holds true. Then the operators $Q_\pm$ are invertible. Let $L_\pm$ be their inverses. Then, in the splitting* (4.9), *the HUM control operator Λ is the pseudodifferential operator of order* 0 *equal to*

$$\Lambda = \Pi_+ \begin{pmatrix} 2L_+ & 0 \\ 0 & 2L_- \end{pmatrix} \Pi_+ + R, \tag{4.18}$$

where R is smoothing. In particular, Λ is an isomorphism of $H^s(M) \oplus H^s(M)$ for all $s \in \mathbb{R}$, and one has the equality of wave-front sets

$$WF^s(\Lambda f) = WF^s(f).$$

Proof. Since we know that M_T is invertible, by (4.13), we just have to verify that the operators $Q_\pm$ are invertible and the operators $\mathcal{T}$ and $\mathcal{T}^*$ are smoothing. Since $Q_\pm$ are elliptic self-adjoint pseudodifferential operators, we just have to verify that $Q_\pm h = 0$ implies $h = 0$. Using (GCC), we get that (2.1) holds true, and we are reduced to show that if $h \in L^2_+$ is such that $\Pi_+ \chi_0 e^{-is\underline{\lambda}} h|_{]0,T[} = 0$, then $h = 0$. Since $\Pi_+(f(x)) = 0$ is equivalent to $f(x)$ independent of x, there exists a function $g(s)$ such that $\chi_0(x) e^{-is\underline{\lambda}} h|_{]0,T[} = g(s)$. If there exists x_0 such that $\chi_0(x_0) = 0$, then $g = 0$ and we get $e^{-is\underline{\lambda}} h|_{(]0,T[\times\omega)} = 0$, and therefore $h = 0$ since (GCC) holds true. Otherwise, χ_0 never vanishes, and $\frac{g(s)}{\chi_0(x)}$ is a solution of the wave equation. This is possible only if $\chi_0(x) = 1/\phi(x)$ with ϕ an eigenvector of Δ, and, since χ_0 is smooth, this implies $\chi_0 = C^{te}$; thus $e^{-is\underline{\lambda}} h$ is independent of x, and $h = 0$ follows from $h \in L^2_+$.

It remains to verify that the operators $\mathcal{T}$ and $\mathcal{T}^*$ are smoothing. Since χ depends on t and is flat at $t = 0$, T, an integration by parts gives

$$\mathcal{T} = \int_0^T e^{is\underline{\lambda}} (\underline{\chi}(T - s, x))^2 e^{is\underline{\lambda}} ds,$$

$$2i\underline{\lambda}\mathcal{T} = -\int_0^T e^{is\underline{\lambda}} \left(\partial_s (\underline{\chi})^2 - [i\underline{\lambda}, (\underline{\chi})^2] \right) e^{is\underline{\lambda}} ds. \tag{4.19}$$

Since $A(s,x,D_x) = \big(\partial_s(\underline{\chi})^2 - [i\underline{\Delta},(\underline{\chi})^2]\big)$ is a smooth family in s of zero order of pseudodifferential operators on M, flat at $t = 0, T$, we can iterate this integration by parts, and this shows that $\mathcal{T}$ is smoothing. □

4.2. The boundary calculus. In this section we first look at the action of the multiplier $\chi_0^2(x)$ on the spaces $H^s(\Omega,\Delta)$. We then study some aspects of the Littlewood–Paley theory at the boundary. Finally, we deduce from this study some useful lemmas.

To begin, we introduce the regularity index $s_0 \in [1/2,+\infty[$, associated with the multiplier χ_0^2. It is chosen so that

$$\tag{4.20}\begin{aligned}&\chi_0^2 \text{ maps } H^s(\Omega,\Delta) \text{ in } H^s(\Omega,\Delta) \quad \forall s \in [0, s_0+2[,\\ &\text{the commutator } [\Delta,\chi_0^2] \text{ maps } H^{s+1}(\Omega,\Delta) \text{ in } H^s(\Omega,\Delta) \quad \forall s \in [0, s_0+2[.\end{aligned}$$

One can take $s_0 = \infty$ if χ_0 is constant near any connected component of the boundary $\partial\Omega$. For arbitrary χ_0, we take $s_0 = 1/2$, and if χ_0 satisfies the additional condition $\partial_n\chi_0|_{\partial\Omega} = 0$, then we take $s_0 = 5/2$.

Since Ω is regular, $F = \partial\Omega$ is a smooth subvariety of $\mathbb{R}^d$, and in geodesic normal coordinates $(x,y) \in \mathbb{R}\times F$ one has near the boundary $\Omega =]0,r[\times F$ for some $r > 0$. Moreover, in these coordinates, the metric is given by

$$\begin{pmatrix} 1 & 0 \\ 0 & g(x,y) \end{pmatrix}$$

and the Laplace operator $\Delta = (\det g)^{-1/2}\sum\partial_i(g^{i,j}(\det g)^{1/2}\partial_j)$ is equal to

$$\tag{4.21}\Delta = \partial_x^2 + a(x,y)\partial_x + R_g(x,y,\partial_y)$$

with $a = \partial_x \log(\det g)^{1/2}$, and $R_g(x_0,y,\partial_y)$ is equal to the Laplace operator on the subvariety $x = x_0$. Thus, with $e = (\det g)^{-1/4} > 0$, one has $\Delta' = e^{-1}\Delta e = \partial_x^2 + R(x,y,\partial_y)$, where R and R_g have the same principal symbol. Therefore, with $u = ev$, one has $\Delta^j u|_{x=0} = 0$ iff $\Delta'^j v|_{x=0} = 0$, and one gets near the boundary

$$\tag{4.22}\begin{aligned}H^s(\Omega,\Delta) &= \{v \in H^s(\Omega),\ \ v|_{x=0} = 0\} \quad \text{for } 1/2 < s < 5/2,\\ H^s(\Omega,\Delta) &= \{v \in H^s(\Omega),\ \ v|_{x=0} = \partial_x^2 v|_{x=0} = 0\} \quad \text{for } 5/2 < s < 9/2,\\ H^s(\Omega,\Delta) &= \{v \in H^s(\Omega),\ \ v|_{x=0} = \partial_x^2 v|_{x=0}\partial_x^4 v|_{x=0} + 2(\partial_x R)\partial_x v|_{x=0} = 0\}\\ &\qquad\qquad \text{for } 9/2 < s < 13/2.\end{aligned}$$

Since with $\chi_2 = \chi_0^2$, $\chi_2 u = e\chi_2 v$, one gets easily from (4.22) that χ_2 maps $H^s(\Omega,\Delta)$ into itself for all $s \in [0,5/2[$, and, under the additional condition $\partial_x\chi_2(0,y) = 0$, χ_2 maps $H^s(\Omega,\Delta)$ into itself for all $s \in [0,9/2[$. Observe that it is impossible in general to get a better result: for $9/2 < s < 13/2$, and with $\partial_x\chi_2(0,y) = 0$, by the third line of (4.22), χ_2 maps $H^s(\Omega,\Delta)$ into itself iff one has

$$\tag{4.23}\begin{aligned}&N(\partial_x v(0,y)) = 0 \quad \forall v \in H^s(\Omega,\Delta),\\ &N(y,\partial_y) = 4\partial_x^3\chi_2(0,y) + 2[\partial_x R(0,y,\partial_y),\chi_2(0,y)].\end{aligned}$$

Obviously, this implies that the first order operator $N(y, \partial_y)$ on the boundary is identically 0. But its principal symbol is proportional to the vector field $\sum k^{i,j}(y)\partial_{y_j}\chi_2(0,y)\partial_{y_i}$, where $k = \partial_x g^{-1}|_{x=0}$. In general, the second fundamental form k is for almost all $y \in \partial\Omega$ nondegenerate, so this implies that $\chi_2(0,y)$ is constant on any connected component of the boundary.

Let us now look at the commutator $[\Delta, \chi_2]$. One has

$$[\Delta', \chi_2] = 2\partial_x\chi_2\partial_x + \partial_x^2\chi_2 + [R, \chi_2]. \tag{4.24}$$

From (4.24) and (4.22), one gets easily that $[\Delta, \chi_2]$ maps $H^{s+1}(\Omega, \Delta)$ into $H^s(\Omega, \Delta)$ for all $s \in [0, 1/2[$, and, under the additional condition $\partial_x\chi_2(0,y) = 0$, χ_2 maps $H^{s+1}(\Omega, \Delta)$ into $H^s(\Omega, \Delta)$ for all $s \in [0, 5/2[$. As above, it is impossible to have a better result in general, except if χ_2 satisfies a strong global geometric hypothesis, for example, if one assumes that χ_2 is constant near any connected component of the boundary.

We shall now study the Littlewood–Paley theory at the boundary. For $z \in \mathbb{C} \setminus [0, \infty[$, we denote by R_z the resolvent of the Laplace operator with Dirichlet boundary condition

$$\begin{aligned} &R_z(g) = f \quad \text{iff} \\ &(z + \Delta)f = g \quad \text{and} \quad f|_{\partial\Omega} = 0. \end{aligned} \tag{4.25}$$

Then, for any $s > -3/2$, R_z maps $H^s(\Omega)$ into $H^{s+2}(\Omega)$. Of course, one has

$$R_z\left(\sum_j a_j e_j\right) = \sum_j \frac{a_j}{z - \omega_j^2} e_j. \tag{4.26}$$

Therefore, for $j = 0, 1, 2$, if Q is a compact subset of $\mathbb{C}$, then there exists C_j such that R_z satisfies the estimates

$$\|h^{-2}R_{zh^{-2}}(f)\|_{H^j(\Omega,\Delta)} \le \frac{C_j h^{-j}}{|\mathrm{Im}(z)|}\|f\|_{L^2} \quad \forall h \in]0,1], \forall z \in Q. \tag{4.27}$$

We denote by K_z the Poisson operator

$$\begin{aligned} &K_z(g) = f \quad \text{iff} \\ &(z + \Delta)f = 0 \quad \text{and} \quad f|_{\partial\Omega} = g. \end{aligned} \tag{4.28}$$

Then, for any s, K_z maps $H^s(\partial\Omega)$ into $H^{s+1/2}(\Omega)$. Finally, we denote by γ_0 the trace operator $\gamma_0(f) = f|_{\partial\Omega}$; for any $s > 1/2$, γ_0 maps $H^s(\Omega)$ into $H^{s-1/2}(\partial\Omega)$. Obviously, one has

$$\begin{aligned} &(z + \Delta)R_z(g) = g \quad \forall g \in H^s(\Omega), \quad s > -3/2, \\ &R_z(z + \Delta)(f) = f - K_z\gamma_0(f) \quad \forall f \in H^s(\Omega), \quad s > 1/2. \end{aligned} \tag{4.29}$$

In particular, one must take care of the fact that Δ and R_z do not commute, since from (4.29) one has

$$[R_z, \Delta] = -K_z\gamma_0. \tag{4.30}$$

More generally, let $A = A(x, \partial_x)$ be a differential operator with smooth coefficients on $\overline{\Omega}$. Then the following commutation relation holds true:

$$[R_z, A] = R_z[A, \Delta]R_z - K_z\gamma_0 AR_z. \tag{4.31}$$

This follows from the fact that for $[R_z, A]f = g$ one has $(z+\Delta)g = Af-(z+\Delta)AR_zf$, $g|_{\partial\Omega} = -\gamma_0 AR_zf$, and $(z + \Delta)AR_zf = [\Delta, A]R_zf + Af$.

Let $\theta \in C_0^\infty(\mathbb{R})$, $h \in]0, 1]$, and let $\theta(h^2|\Delta|)$ be the operator

$$\theta(h^2|\Delta|)\Big(\sum_j a_je_j\Big) = \sum_j \theta(h^2\omega_j^2)a_je_j. \tag{4.32}$$

Then $\theta(h^2|\Delta|)$ maps $H^s(\Omega, \Delta)$ into $H^\infty(\Omega, \Delta)$ for all $s \in \mathbb{R}$. Let $\widetilde{\theta} \in C_0^\infty(\mathbb{C})$ be an almost analytic extension of θ. This means that $\widetilde{\theta}(x) = \theta(x)$ for all $x \in \mathbb{R}$ and $\overline{\partial}\widetilde{\theta}(z) \in O(|\mathrm{Im}(z)|^\infty)$. Almost analytic extensions have been introduced in a lecture seminar by Hörmander in [Hör68], and one can take

$$\widetilde{\theta}(x + iy) = \sum_{k=0}^{\infty} \frac{\theta^{(k)}(x)}{k!}(iy)^k\chi(a_ky) \tag{4.33}$$

with $\chi \in C_0^\infty(\mathbb{R})$ equal to 1 near 0 and the sequence a_k tending to $+\infty$ sufficiently fast when $k \to \infty$. From $\overline{\partial}(\frac{1}{\pi z}) = \delta_0$, one gets that the following formula, which is called the Helffer–Sjöstrand formula, holds true:

$$\theta(h^2|\Delta|) = \frac{-1}{\pi}\int_{\mathbb{C}} \frac{\overline{\partial}\widetilde{\theta}(z)}{z - h^2|\Delta|}L(dz) = \frac{-h^{-2}}{\pi}\int_{\mathbb{C}} \overline{\partial}\widetilde{\theta}(z)R_{h^{-2}z}L(dz), \tag{4.34}$$

where $L(dz)$ is the Lebesgue measure on $\mathbb{C}$. For $j = 0, 1$, let A_j be a differential operator of order j such that $\gamma_0 f = 0$ implies $\gamma_0 A_jf = 0$. This is always true if $j = 0$, and if $j = 1$, then it holds if the coefficient of the normal derivative ∂_n in A_1 vanishes at the boundary. Then we get from (4.31), since $\gamma_0 A_jR_z = 0$,

$$[\theta(h^2|\Delta|), A_j] = \frac{h^2}{\pi}\int_{\mathbb{C}} \overline{\partial}\widetilde{\theta}(z)\frac{1}{z + h^2\Delta}[\Delta, A_j]\frac{1}{z + h^2\Delta}L(dz). \tag{4.35}$$

Since $[\Delta, A_j]$ is a differential operator of order $j + 1 \leq 2$, we get from (4.27)

$$\begin{aligned} &\left\|[\Delta, A_j]\frac{1}{z + h^2\Delta}(f)\right\|_{L^2} \leq \frac{C_jh^{-j-1}}{|\mathrm{Im}(z)|}\|f\|_{L^2}, \\ &\left\|\frac{1}{z + h^2\Delta}[\Delta, A_j]\frac{1}{z + h^2\Delta}(f)\right\|_{L^2} \leq \frac{C_jh^{-j-1}}{|\mathrm{Im}(z)|^2}\|f\|_{L^2}. \end{aligned} \tag{4.36}$$

Therefore, one gets from (4.35) and (4.36) the estimate

$$\|[\theta(h^2|\Delta|), A_j]\|_{L^2} \leq C_jh^{1-j}. \tag{4.37}$$

Let us now study the operator Q_+, where $(\chi)^2(t, x) = \psi^2(T - t)\chi_0^2(x)$:

$$Q_+ = \int_0^T e^{it\lambda}(\chi)^2e^{-it\lambda}dt. \tag{4.38}$$

By the definition of the regularity index s_0 given in (4.20), the self-adjoint operator Q_+ obviously maps $H^s(\Omega,\Delta)$ into itself for all $s \in\,]-s_0-2, s_0+2[$. The following lemma shows that the integration in time in (4.38) induces a better result.

Lemma 4.2. *The operator Q_+ is bounded on $H^s(\Omega,\Delta)$ for all s.*

Proof. Since Q_+ is self-adjoint on L^2 we may assume $s \geq 0$. We set

$$Q_+ = \sum_{i,j} \mathcal{C}_{i,j}, \quad \mathcal{C}_{i,j} = \int_0^T e^{it\lambda}\psi_i(D)(\chi)^2\psi_j(D)e^{-it\lambda}dt. \tag{4.39}$$

We have to check that the matrix $2^{is}\|\mathcal{C}_{i,j}\|_{L^2}2^{-js}$ is bounded on $l^2(\mathbb{N})$. Take $M\geq 2$. For $|i-j|\leq M$ one obviously has $\|\mathcal{C}_{i,j}\|_{L^2}\leq C$, so the near diagonal terms contribute to an operator which is bounded on $H^s(\Delta)$ for all s. For the case $i > j+M$, by integration by parts in (4.39) we get

$$\mathcal{C}_{i,j} = \int_0^T e^{it\lambda}\lambda^{-1}\psi_i(D)\Big(i\partial_t(\chi)^2+(\chi)^2\lambda\Big)\psi_j(D)e^{-it\lambda}dt.$$

Using N iterations of this integration by parts, we get for $i > M+j$

$$\|\mathcal{C}_{i,j}\|_{L^2}\leq C_N 2^{-N(i-j)}. \tag{4.40}$$

Since the adjoint on L^2 of $\mathcal{C}_{i,j}$ is $\mathcal{C}_{j,i}$, we get for $|i-j|>M$

$$\|\mathcal{C}_{i,j}\|_{L^2}\leq C_N 2^{-N|i-j|}. \tag{4.41}$$

Thus we obtain that for all N there exists C_N such that for all i,j

$$2^{is}\|\mathcal{C}_{i,j}\|_{L^2}2^{-js}\leq C_N 2^{-(N-s)|i-j|}. \tag{4.42}$$

This matrix is bounded on l^2 for N large, and thus the self-adjoint operator Q_+ is bounded on $H^s(\Omega,\Delta)$ for all s. The proof of our lemma is complete. □

Lemma 4.3. *The operator $[Q_+,\lambda^{-s}]$ is bounded from L^2 into $H^{s+1}(\Omega,\Delta)$ for any $s\in[0,2[$. Moreover, the operator $[Q_+,\lambda]$ is bounded on L^2.*

Proof. One has

$$[Q_+,\lambda^{-s}] = \sum_{i,j}\mathcal{D}_{i,j}, \quad \mathcal{D}_{i,j} = \int_0^T e^{it\lambda}\psi_i(D)[(\chi)^2,\lambda^{-s}]\psi_j(D)e^{-it\lambda}dt. \tag{4.43}$$

Obviously, for each i,j, the operator $\mathcal{D}_{i,j}$ is bounded on L^2, and, by the Littlewood–Paley theory, we have to check that for $s\in[0,2[$ the matrix

$$2^{i(s+1)}\|\mathcal{D}_{i,j}\|_{L^2} \tag{4.44}$$

is bounded on $l^2(\mathbb{N})$. We shall prove that for all $N\geq 0$ there exists C_N such that

$$2^{i(s+1)}\|\mathcal{D}_{i,j}\|_{L^2}\leq C_N 2^{-N|i-j|} \quad \forall i,j. \tag{4.45}$$

We fix $M\geq 2$ and study separately the two cases $|i-j|\leq M$ (near diagonal terms) and $|i-j|>M$ (off-diagonal terms). Let us first look at the case $|i-j|\leq M$. Let γ be a contour in $Re(z)>0$ which is the union of two half lines with end point $\omega_1^2/2$ and with angle $\pm\pi/4$ with the real axis. We take the up-down orientation on γ. Then, by the Cauchy formula, one has

$$\lambda^{-s} = \frac{1}{2i\pi}\int_\gamma \frac{z^{-s/2}}{z+\Delta}dz. \tag{4.46}$$

Thus we get from (4.31), since $\gamma_0(\chi)^2 R_z = 0$,

$$[(\chi)^2, \lambda^{-s}] = \frac{1}{2i\pi}\int_\gamma \frac{z^{-s/2}}{z+\Delta}[\Delta,(\chi)^2]\frac{1}{z+\Delta}dz. \tag{4.47}$$

Let $A_1 = [\Delta, (\chi)^2]$. Then from (4.47) we get

$$\mathcal{D}_{i,j} = \int_0^T e^{it\lambda}\Big(\frac{1}{2i\pi}\int_\gamma z^{-s/2}\frac{\psi_i(D)}{z+\Delta}A_1\frac{\psi_j(D)}{z+\Delta}dz\Big)e^{-it\lambda}dt. \tag{4.48}$$

For $z \in \gamma$, one has $\|\frac{\psi_i(D)}{z+\Delta}\|_{L^2} \le \frac{C}{|z|+2^{2i}}$, and, since A_1 is a first order operator,

$$\Big\|A_1\frac{\psi_j(D)}{z+\Delta}f\Big\|_{L^2} \le C\Big\|\frac{\psi_j(D)}{z+\Delta}f\Big\|_{H^1(\Omega,\Delta)} \le C\frac{2^j}{|z|+2^{2j}}\|f\|_{L^2}.$$

Thus we get that (4.45) holds true from the estimate for $s<2$:

$$\int_\gamma |z|^{-s/2}\frac{2^k}{(|z|+2^{2k})^2}|dz| \le C2^{-k(1+s)}. \tag{4.49}$$

For $|i-j|>M$ we write $\mathcal{D}_{i,j} = \mathcal{D}^1_{i,j} - \mathcal{D}^2_{i,j}$ with

$$\begin{aligned}\mathcal{D}^2_{i,j} &= \int_0^T e^{it\lambda}\lambda^{-s}\psi_i(D)(\chi)^2\psi_j(D)e^{-it\lambda}dt,\\ \mathcal{D}^1_{i,j} &= \int_0^T e^{it\lambda}\psi_i(D)(\chi)^2\psi_j(D)\lambda^{-s}e^{-it\lambda}dt,\end{aligned} \tag{4.50}$$

and we bound each of these two terms.

Since $|i-j|\ge M$, one has, with $\widetilde{\psi}\in C_0^\infty(]1/8,2[)$ equal to 1 near the support of ψ, $\psi_i(D)(\chi)^2\psi_j(D) = \psi_i(D)[(\chi)^2,\psi_j(D)]\widetilde{\psi}_j(D)\widetilde{\psi}_i(D)[\psi_i(D),(\chi)^2]\psi_j(D)$. Thus, for $i\ge j+M$, one has for all $N\ge 0$

$$\begin{aligned}\mathcal{D}^2_{i,j} &= \int_0^T e^{it\lambda}\lambda^{-s-N}\widetilde{\psi}_i(D)(i\partial_t)^N\Big([\psi_i(D),(\chi)^2]\psi_j(D)e^{-it\lambda}\Big)dt,\\ \mathcal{D}^1_{i,j} &= \int_0^T e^{it\lambda}\lambda^{-N}\widetilde{\psi}_i(D)(i\partial_t)^N\Big([\psi_i(D),(\chi)^2]\psi_j(D)\lambda^{-s}e^{-it\lambda}\Big)dt,\end{aligned} \tag{4.51}$$

and from (4.37) we get $2^{i(s+1)}\|\mathcal{D}^2_{i,j}\|_{L^2} \le C_N 2^{-N(i-j)}$, and $2^{i(s+1)}\|\mathcal{D}^2_{i,j}\|_{L^2} \le C_N$ $2^{-(N-s)(i-j)}$. Thus we get that (4.45) holds true in that case. For $j\ge i+M$, we write

$$\begin{aligned}\mathcal{D}^2_{i,j} &= \int_0^T (-i\partial_t)^N\Big(e^{it\lambda}\lambda^{-s}\psi_i(D)[(\chi)^2,\psi_j(D)]\Big)\widetilde{\psi}_j(D)\lambda^{-N}e^{-it\lambda}dt,\\ \mathcal{D}^1_{i,j} &= \int_0^T (-i\partial_t)^N\Big(e^{it\lambda}\psi_i(D)[(\chi)^2,\psi_j(D)]\Big)\widetilde{\psi}_j(D)\lambda^{-s-N}e^{-it\lambda}dt,\end{aligned} \tag{4.52}$$

and from (4.37) we get $2^{i(s+1)}\|\mathcal{D}^2_{i,j}\|_{L^2} \leq C_N 2^{-(N+1)(j-i)}$, and $2^{i(s+1)}\|\mathcal{D}^2_{i,j}\|_{L^2} \leq C_N 2^{-(N+s+1)(j-i)}$.

Thus we get that (4.45) holds true. Finally, one has $[Q_+, \lambda] = \lambda[\lambda^{-1}, Q_+]\lambda$, and, since $B = [\lambda^{-1}, Q_+]$ is skew-adjoint, it follows from the previous result that B is bounded from L^2 into $H^2(\Omega, \Delta)$ and from $H^{-2}(\Omega, \Delta)$ into L^2. Hence B is bounded from $H^{-1}(\Omega, \Delta)$ into $H^1(\Omega, \Delta)$. The proof of our lemma is complete. □

4.3. Strichartz inequalities. Here we recall the Strichartz inequalities for the linear wave equation, which play a key role in section 3. The interested reader can find them, for example, in [Str77], [GV85], [Gér96], in [Kap94] for compact manifolds without boundary, and in [BLP07] for the open subdomain of $\mathbb{R}^3$. Consider the linear wave equation on a three-dimensional compact manifold M without boundary:

$$\left\{\begin{array}{l} \Box u = F \in L^1([0, +\infty[, L^2(M)), \\ (u(0), \partial_t u(0)) \in H^1(M) \times L^2(M). \end{array}\right. \tag{4.53}$$

Then the usual Strichartz inequality reads as follows.

Lemma 4.4. *Let $r \in [2, +\infty[$ and q be given by $1/q + 1/r = 1/2$. For every $T > 0$, there exists $C = C(T) > 0$ such that for every solution u of* (4.53) *one has*

(4.54)
$$\|u\|_{L^q([0,T],L^{3r}(M))} \leq C\left[\|F\|_{L^1([0,T],L^2(M))} + \|\partial_t u(0)\|_{L^2(M)} + \|\nabla_x u(0)\|_{L^2(M)}\right].$$

Now concerning the wave equation on a bounded open subset of $\mathbb{R}^3$, with smooth boundary, the situation is more delicate. Such estimates as well as the global well posedness were not known before the very recent work [BLP07] by Burq, the second author, and Planchon. More precisely, they prove the following.

Lemma 4.5. *For every $T > 0$, there exists $C > 0$ such that every solution u of*

$$\left\{\begin{array}{l} \Box u = F \quad \text{in} \quad]0, +\infty[\times\Omega, \\ u_{|\partial\Omega} = 0, \quad (u(0), \partial_t u(0)) = (u_0, u_1) \end{array}\right. \tag{4.55}$$

satisfies

$$\begin{aligned} &\|u\|_{L^5(0,T;W_0^{\frac{3}{10},5}(\Omega))} + \|u\|_{C^0(0,T;H^1_0(\Omega))} + \|\partial_t u\|_{C^0(0,T;L^2(\Omega))} \\ &\leq C(\|u_0\|_{H^1_0(\Omega)} + \|u_1\|_{L^2(\Omega)} + \|F\|_{L^1(0,T;L^2(\Omega))}). \end{aligned} \tag{4.56}$$

4.4. Regularity of the composition. In this section we study the regularity of the composition $f(u)$ representing the semilinear term of (1.40). We give an analogous result to Theorem 8 of [DLZ03], with a slightly different proof.

For a tempered distribution u on $\mathbb{R}^3$, we denote by $(u_q)_{q\geq -1}$ its dyadic decomposition: $u = u_{-1} + \sum_{q\geq 0} u_q$. It is the usual Littlewood–Paley decomposition. The interested reader can find a good exposition of it, for example, in [AG91], [Mey90], or [Che95]. Here we give an abstract multiplier lemma that will be the basis of the composition theorem below.

Lemma 4.6 (Meyer's multipliers). *Let $\alpha \in]5/4, 5/3]$, and let $(m_q)_{\{q\geq -1\}}$ be a sequence of C^∞ functions such that $\sum_{|\mu|=l} \|\partial^\mu m_q\|_{L^{2\alpha}} \leq C_l 2^{ql}$ for $l = 0, 1, 2$. Then the operator*

$$M : u = \sum_{q\geq -1} u_q \longmapsto Mu = \sum_{q\geq -1} m_q u_q \tag{4.57}$$

is continuous from $W^{\frac{3}{10},5}(R^3)$ *into* $H^\sigma(R^3)$*, where* $\sigma = \frac{3}{10\alpha}(4\alpha - 5)$*. More precisely,*

$$\|Mu\|_{H^\sigma(\mathbb{R}^3)} \le C\,\|u\|_{W^{\frac{3}{10},5}(\mathbb{R}^3)} \quad \textit{with} \quad C \le \text{Const.}\sum_{l\le 2} C_l. \tag{4.58}$$

Proof. The proof is similar word by word to the proof of Lemma 7 in [DLZ03]. We sketch it for the convenience of the reader. The principal material is borrowed from [AG91, Lemma 2.2].

Let $u = u_{-1}+\sum_{q\ge 0} u_q$ be the Littlewood–Paley decomposition of u. The spectrum of u_q is contained in the ring $2^{q-1} \le |\xi| \le 2^{q+1}$. On the other hand, we decompose m_q as $m_q = m_{q,-1} + \sum_{k\ge 0} m_{q,k}$, where the spectrum of $m_{q,-1}$ is contained in a ball of radius 2^q, and those of $m_{q,k}$ for $k \ge 0$ are contained in rings of order 2^{q+k}. We set $M_k u = \sum_{q\ge -1} m_{q,k} u_q$, $k \ge -1$. We will show that each M_k is continuous from $W^{\frac{3}{10},5}(\mathbb{R}^3)$ to $H^\sigma(\mathbb{R}^3)$ and that the corresponding operator series converges normally.

For $k \ge 0$, the terms in $M_k u$ have their spectra in annulae of the order of 2^{q+k}. We have

$$||m_{q,k}u_q||_{L^2} \le ||m_{q,k}||_{L^{2\alpha}}||u_q||_{L^{2\beta}}$$

with $1/\alpha + 1/\beta = 1$. Thanks to the imbedding $W^{\frac{3}{10},5}(\mathbb{R}^3) \hookrightarrow W^{\sigma,2\beta}(\mathbb{R}^3)$ with

$$\sigma = \frac{3}{10\beta}(5-\beta) = \frac{3}{10\alpha}(4\alpha - 5) \in]0, 3/10]$$

we estimate

$$||u_q||_{L^{2\beta}} \le c_q 2^{-\sigma q}||u||_{W^{\sigma,2\beta}} \le c_q 2^{-\sigma q}||u||_{W^{\frac{3}{10},5}},$$

where $\sum c_q^2 < \infty$. Moreover,

$$\|m_{q,k}\|_{L^{2\alpha}} \le C_l 2^{-2k}. \tag{4.59}$$

This is true, indeed, since, by hypothesis,

$$\|m_{q,k}\|_{L^{2\alpha}} \le C \sum_{|\mu|=2} \|\partial^\mu m_q\|_{L^{2\alpha}}\, 2^{-2(q+k)} \le C_l 2^{-2k}.$$

Thus

$$\|m_{q,k}u_q\|_{L^2} \le C_l 2^{-2k} c_q 2^{-\sigma q}||u||_{W^{\frac{3}{10},5}} \le C_l 2^{-k(2-\sigma)} c_q 2^{-(q+k)\sigma}||u||_{W^{\frac{3}{10},5}}.$$

Applying then the synthesis lemma (Lemma 2.1) in [AG91], we deduce that M_k is continuous from $W^{\frac{3}{10},5}(\mathbb{R}^3)$ to $H^\sigma(\mathbb{R}^3)$, with a norm of the order of $C_l 2^{-k(2-\sigma)}$. The terms in $M_{-1}u$ are treated in a similar way and give the same estimate. Finally, it is obvious that the operator series $M = \sum M_k$ converges normally in the space of continuous operators from $W^{\frac{3}{10},5}(\mathbb{R}^3)$ to $H^\sigma(\mathbb{R}^3)$ and satisfies estimate (4.58). □

Now we study the regularity of the function $f(u)$.

THEOREM 4.7. *Let* $p \in [4, 5[$*. Let* v *be a function in* $L^5([0,T], W^{\frac{3}{10},5}(R^3))$ *and* f *a function satisfying conditions* (1.39) *and* (1.40)*. Then* $f(v) \in L^1([0,T], H^r(R^3))$*, with* $r = \frac{3}{10}(5-p)$*.*

Remark 4.8. This theorem states the same result as Theorem 8 in [DLZ03], except that, in the present work, only the Strichartz norms $\|v\|_{L^q(0,T;L^{3r})}$ with $q \geq 5$ and $1/q + 1/r = 1/2$ are supposed to be finite.

Proof. As in [DLZ03], we follow [Mey90] and [AG91] and write

$$f(v) = f(S_0 v) + f(S_1 v) - f(S_0 v) + \cdots + f(S_{q+1} v) - f(S_q v) + \cdots ,$$

where

$$S_q v = v_{-1} + v_0 + \cdots + v_q.$$

First, we will work with a fixed s in $[0, T]$. Then we will examine the integrability in time of the H^r-norms. The first term $f(S_0 v)$ has the regularity of f, so it is easy to treat. For $q \geq 0$, we write $f(S_{q+1}v) - f(S_q v) = m_q v_q$, with $m_q = \int_0^1 f^{'}(S_q v + t v_q)dt$, and we show that the m_q's are Meyer's multipliers in the sense of the previous lemma. More precisely, following line by line the proof of [DLZ03], we establish the estimate

$$\sum_{|\mu|=l} \|\partial^\mu m_q\|_{L^{2\alpha}} \leq C(1 + \|v\|_{L^{10}}^{p-1}) 2^{ql} \quad \text{for } l \leq 2 \tag{4.60}$$

with $\alpha = \frac{5}{p-1} \in]5/4, 5/3]$. This implies

$$\begin{aligned}\int_0^T \|f(v)\|_{H^r}\, ds &\leq C \int_0^T (1 + \|v(s)\|_{L^{10}}^{p-1}) \, \|v(s)\|_{W^{\frac{3}{10},5}}\, ds \\ &\leq C \int_0^T (1 + \|v(s)\|_{W^{\frac{3}{10},5}}^{p-1}) \, \|v(s)\|_{W^{\frac{3}{10},5}}\, ds \\ &\leq C \int_0^T (\|v(s)\|_{W^{\frac{3}{10},5}} + \|v(s)\|_{W^{\frac{3}{10},5}}^{p}) ds,\end{aligned}$$

and this is finite since $p < 5$. □

REFERENCES

[AG91] S. Alinhac and P. Gérard, *Opérateurs pseudo-différentiels et théorème de Nash-Moser*, Savoirs actuels, InterEditions, Paris, Editions du CNRS, Meudon, 1991.

[BLR92] C. Bardos, G. Lebeau, and J. Rauch, *Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary*, SIAM J. Control Optim., 30 (1992), pp. 1024–1065.

[BG97] N. Burq and P. Gérard, *Condition nécessaire et suffisante pour la contrôlabilité exacte de l'équation des ondes*, C. R. Acad. Sci. Paris Sér. I Math., 325 (1997), pp. 749–752.

[BLP07] N. Burq, G. Lebeau, and F. Planchon, *Global existence for energy critical waves in 3-D domains*, J. Amer. Math. Soc., 21 (2008), pp. 831–845.

[Che95] J.-Y. Chemin, *Fluides parfaits incompressibles*, Astérisque 209, SMF, Paris, 1995.

[DLZ03] B. Dehman, G. Lebeau, and E. Zuazua, *Stabilization and control for the subcritical semilinear wave equation*, Ann. Sci. École Norm. Sup. (4), 36 (2003), pp. 525–551.

[Gér91] P. Gérard, *Microlocal defect measures*, Comm. Partial Differential Equations, 16 (1991), pp. 1762–1794.

[Gér96] P. Gérard, *Oscillation and concentration effects in semilinear dispersive wave equations*, J. Funct. Anal., 41 (1996), pp. 60–98.

[GV85] J. Ginibre and G. Velo, *The global Cauchy problem for the nonlinear Klein-Gordon equation*, Math. Z., 189 (1985), pp. 487–505.

[Hör68] L. Hörmander, *Seminar*, Lectures Notes at the Nordic Summer School of Mathematics, 1968.

[Hör85] L. Hörmander, *The Analysis of Linear Partial Differential Operators: Pseudodifferential Operators*, III, Grundlehren Math. Wiss. 274, Springer-Verlag, Berlin, 1985.

[Kap94] L. Kapitanski, *Weak and yet weaker solutions of semilinear wave equations*, Comm. Partial Differential Equations, 19 (1994), pp. 1629–1676.

[Leb92] G. Lebeau, *Contrôle de l'équation de Schrödinger*, J. Math. Pures Appl. (9), 71 (1992), pp. 267–291.

[LN08] G. Lebeau and M. Nodet, *Experimental Study of the HUM Control Operator for Linear Waves*, http://hal.inria.fr/ (2008).

[Lio88] J.-L. Lions, *Contrôlabilité exacte, perturbations et stabilisation de systèmes distribués. Tome* 1, Rech. Math. Appl. 8, Masson, Paris, 1988.

[Mey90] Y. Meyer, *Ondelettes et opérateurs*, Actualites Math., Hermann, Paris, 1990.

[Rus78] D. L. Russell, *Controllability and stabilizability theory for linear partial differential equations: Recent progress and open questions*, SIAM Rev., 20 (1978), pp. 639–739.

[Str77] R. Strichartz, *Restriction of Fourier transform to quadratic surfaces and decay of solutions of the wave equation*, Duke Math. J., 44 (1977), pp. 705–714.

[Zua90] E. Zuazua, *Exact controllability for the semilinear wave equation*, J. Math. Pures Appl. (9), 69 (1990), pp. 1–31.

[Zua93] E. Zuazua, *Exact controllability for semilinear wave equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 10 (1993), pp. 109–129.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 2, pp. 551–572

© 2009 Society for Industrial and Applied Mathematics

A PROOF OF THE SMOOTHNESS OF THE FINITE TIME HORIZON AMERICAN PUT OPTION FOR JUMP DIFFUSIONS[*]

ERHAN BAYRAKTAR[†]

Abstract. We give a new proof of the fact that the value function of the finite time horizon American put option for a jump diffusion, when the jumps are from a compound Poisson process, is the classical solution of a free boundary equation. We also show that the value function is C^1 across the optimal stopping boundary. Our proof, which uses only the classical theory of parabolic partial differential equations of [A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice–Hall, Englewood Cliffs, NJ, 1964] and [A. Friedman, *Stochastic Differential Equations and Applications*, Dover, Mineola, NY, 2006], is an alternative to the proof that uses the theory of viscosity solutions (see [H. Pham, *Appl. Math. Optim.*, 35 (1997), pp. 145–164]). This new proof relies on constructing a monotonous sequence of functions, each of which is a value function of an optimal stopping problem for a geometric Brownian motion, converging to the value function of the American put option for the jump diffusion uniformly and exponentially fast. This sequence is constructed by iterating a functional operator that maps a certain class of convex functions to classical solutions of corresponding free boundary equations. On the other hand, since the approximating sequence converges to the value function exponentially fast, it naturally leads to a good numerical scheme.

Key words. optimal stopping, Markov processes, jump diffusions, American options, integro-differential equations, parabolic free boundary equations

AMS subject classifications. 60G40, 62L15, 60J75

DOI. 10.1137/070686494

1. Introduction.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space hosting a Wiener process $W = \{W_t; t \geq 0\}$ and a Poisson random measure N on $\mathbb{R}_+ \times \mathbb{R}_+$, with mean measure $\lambda\nu(dx)dt$ (in which ν is a probability measure on $\mathbb{R}_+$), independent of the Wiener process. We will consider a Markov process $S = \{S_t; t \geq 0\}$ of the form

$$dS_t = \mu S_t dt + \sigma S_t dW_t + S_{t-}\int_{\mathbb{R}_+}(z-1)N(dt,dz). \tag{1.1}$$

In this model, if the stock price jumps at time t, then it moves from S_{t-} to $S_t = ZS_{t-}$, in which Z is a positive random variable whose distribution is given by ν. Note that when $Z < 1$ the stock price jumps down and when $Z > 1$ the stock price jumps up. In the Merton jump diffusion model $Z = \exp(Y)$, in which Y is a Gaussian random variable. We will take $\mu = r + \lambda - \lambda\xi$, in which $\xi = \int_{\mathbb{R}_+} xv(dx) < \infty$, so that $(e^{-rt}S_t)_{t\geq 0}$ is a martingale; i.e., $\mathbb{P}$ is a risk neutral measure. The constant $r \geq 0$ is the interest rate, and the constant $\sigma > 0$ is the volatility. We assume the risk neutral pricing measure $\mathbb{P}$, and hence the parameters of the problem, are fixed as a result of a calibration to historical data. The value function of the American put option pricing problem is

$$V(x,T) := \sup_{\tau\in\widetilde{\mathcal{S}}_{0,T}} \mathbb{E}^x\{e^{-r\tau}(K-S_\tau)^+\}, \tag{1.2}$$

[*] Received by the editors March 27, 2007; accepted for publication (in revised form) September 22, 2008; published electronically February 13, 2009. This research was supported in part by the National Science Foundation.
http://www.siam.org/journals/sicon/48-2/68649.html

[†] Department of Mathematics, University of Michigan, Ann Arbor, MI 48109 (erhan@umich.edu).

in which $\widetilde{\mathcal{S}}_{0,T}$ is the set of stopping times (of the filtration generated by W and N) that take values in $[0,T]$, and $\mathbb{E}^x$ is the expectation under the probability measure $\mathbb{P}$, given that $S_0 = x$.

We will show that V is the classical solution of a free boundary equation and that it satisfies the *smooth fit principle*; i.e., V is continuously differentiable with respect to its first variable at the optimal stopping boundary. We argue these facts by showing that V is the fixed point of an operator, which we will denote by J, that maps a given function to the value function of an optimal stopping problem for a geometric Brownian motion. This operator acts as a regularizer: As soon as the given function f has some certain regularity properties, we show that Jf is the unique classical solution of a corresponding free boundary equation and that it satisfies the smooth fit principle. The proof of the main result concludes once we show that V has these certain regularity properties. In this last step we make use of a sequence (which is constructed by iterating J starting with the pay-off function of the put option) that converges to V uniformly and exponentially fast. Incidentally, this sequence yields a numerical procedure, whose accuracy versus speed characteristics can be controlled. Each element of this sequence is an optimal stopping problem for geometric Brownian motion and can be readily calculated using classical finite difference methods (see, e.g., [18] for the implementation of these methods). An alternative proof of the regularity of V was given in [14]. This proof used a combination of the results in [8] and the theory of viscosity solutions. In particular the proof of Proposition 3.1 in [14] is carried out (details are not provided but hinted) using arguments similar to those used in the proof of Proposition 5.3 in [15]. The latter proof uses the uniqueness results of [9] for viscosity solutions.

The infinite horizon American put options for jump diffusions were analyzed in [3] using the iterative scheme we describe here. The main technical difficulty in the current paper stems from the fact that each element in the approximating sequence solves a parabolic rather than an elliptic problem. In fact, in the infinite horizon case one can obtain a closed form representation for the value function, which is not possible in the finite horizon case. We make use of the results of [8] and Chapter 2 of [10] (also see Chapter 7 of [13]) to study the properties of the approximating sequence. For example, we show that the approximating sequence is bounded with respect to the Hölder seminorm (see page 61 in [7] for a definition), which is used to argue that the limit of the approximating sequence (which is a fixed point of J) solves a corresponding free boundary equation.

Somewhat similar approximation techniques to the one we employ were used to solve optimal stopping problems for *diffusions*; see, e.g., [2] for perpetual optimal stopping problems with nonsmooth pay-off functions and [6], [5] for finite time horizon American put option pricing problems for geometric Brownian motion. On the other hand, [1] and [11] consider the smooth fit principle for the infinite horizon American put option pricing problems for one-dimensional exponential Lévy processes using the fluctuation theory. Also see [4] for the analysis of the smooth fit principle for a multidimensional infinite horizon optimal stopping problem.

The next two sections prepare the proof of our main result, Theorem 3.1, in a sequence of lemmas and corollaries. In the next section, we introduce the functional operator J, which maps a given function to the value function of an optimal stopping problem for a geometric Brownian motion. We then analyze the properties of J. For example, J preserves convexity with respect to the first variable; the increase in the Hölder seminorm after the application of J can be controlled; J maps certain classes of functions to the classical solutions of free boundary equations. In section 3, we construct a sequence of functions that converge to the smallest fixed point of the operator J. We show that the sequence is bounded in the Hölder norm and satisfies certain regularity properties using results of section 2. We eventually arrive at the fact that the smallest fixed point of J is equal to V. As a result the regularity properties of V follow.

2. A functional operator and its properties. Let us define an operator J through its action on a test function $f : \mathbb{R}_+ \times \bar{\mathbb{R}}_+ \to \mathbb{R}_+$: The operator J takes the function f to the value function of the following optimal stopping problem:

$$Jf(x,T) = \sup_{\tau \in \mathcal{S}_{0,T}} \mathbb{E}^x \left\{ \int_0^\tau e^{-(r+\lambda)t} \lambda \cdot Pf(S_t^0, T-t)dt + e^{-(r+\lambda)\tau}(K - S_\tau^0)^+ \right\}, \tag{2.1}$$

in which

$$Pf(x, T-t) = \int_{\mathbb{R}_+} f(xz, T-t)\nu(dz), \quad x \geq 0. \tag{2.2}$$

We will extend $T \to Jf(x,T)$ onto $[0,\infty]$ by letting

$$Jf(x,\infty) = \lim_{T\to\infty} Jf(x,T). \tag{2.3}$$

Here, $S^0 = \{S_t^0; t \geq 0\}$ is the solution of

$$dS_t^0 = \mu S_t^0 dt + \sigma S_t^0 dW_t, \quad S_0^0 = x, \tag{2.4}$$

whose infinitesimal generator is given by

$$\mathcal{A} := \frac{1}{2}\sigma^2 x^2 \frac{d^2}{dx^2} + \mu x \frac{d}{dx}. \tag{2.5}$$

In (2.1), $\mathcal{S}_{[0,T]}$ denotes the set of stopping times of S^0 which take values in $[0,T]$. Note that

$$S_t^0 = xH_t, \tag{2.6}$$

where

$$H_t = \exp\left\{ \left(\mu - \frac{1}{2}\sigma^2\right) t + \sigma W_t \right\}. \tag{2.7}$$

The next remark characterizes the optimal stopping times of (2.1) using the Snell envelope theory.

Remark 2.1. Let us denote

$$Y_t := \int_0^t e^{-(r+\lambda)s} \lambda \cdot Pf(S_t^0, T-s)ds + e^{-(r+\lambda)t}(K - S_t^0)^+. \tag{2.8}$$

Using the strong Markov property of S^0, we can determine the Snell envelope of Y as

$$\xi_t := \sup_{\tau \in \mathcal{S}_{t,T}} \mathbb{E}\left\{Y_\tau | \mathcal{F}_t\right\} \tag{2.9}$$

$$= e^{-(\lambda+r)t}Jf(S_t^0, T-t) + \int_0^t e^{-(r+\lambda)s}\lambda\, Pf(S_s^0, T-s)ds, \quad t \in [0,T].$$

Theorem D.12 in [10] implies that the stopping time

$$(2.10)\ \ \tau_x := \inf\{t \in [0,T) : \xi_t = Y_t\} \wedge T = \inf\{t \in [0,T] : Jf(S_t^0, T-t) = (K-S_t^0)^+\}$$

satisfies

$$(2.11) \quad Jf(x,T) = \mathbb{E}^x \left\{ \int_0^{\tau_x} e^{-(r+\lambda)t}\lambda \cdot Pf(S_t^0, T-t)dt + e^{-(r+\lambda)\tau_x}(K - S_{\tau_x}^0)^+ \right\}.$$

Moreover, the stopped process $(e^{-(r+\lambda)(t\wedge\tau_x)}Jf(S_{t\wedge\tau_x}^0, T-t\wedge\tau_x) + \int_0^{t\wedge\tau_x} e^{-(r+\lambda)s}\lambda \cdot Pf(S_s^0, T-s)ds)_{t\geq 0}$ is a martingale. The second infimum in (2.10) is less than T because $Jf(S_T^0, 0) = (K-S_T^0)^+$.

When f is bounded, it follows from the bounded convergence theorem that (using the results of [3] and arguments similar to the ones used in Corollary 7.3 in Chapter 2 of [10])

$$(2.12)\ \ Jf(x,\infty) = \sup_{\tau\in\mathcal{S}_{0,\infty}} \mathbb{E}^x \left\{ \int_0^{\tau} e^{-(r+\lambda)t}\lambda \cdot Pf(S_t^0, \infty)dt + e^{-(r+\lambda)\tau}(K - S_\tau^0)^+ \right\}.$$

The next three lemmas on the properties of J immediately follow from the definition in (2.1). The first lemma states that J preserves monotonicity.

Lemma 2.1. *Let $T \to f(x,T)$ be nondecreasing and $x \to f(x,T)$ be nonincreasing. Then $T \to Jf(x,T)$ is nondecreasing and $x \to Jf(x,T)$ is nonincreasing.*

The operator J preserves boundedness and order.

Lemma 2.2. *Let $f : \mathbb{R}_+ \times \bar{\mathbb{R}}_+ \to \mathbb{R}_+$ be a bounded function. Then Jf is also bounded. In fact,*

$$(2.13) \qquad 0 \leq \|Jf\|_\infty \leq K + \frac{\lambda}{r+\lambda}\|f\|_\infty.$$

Lemma 2.3. *For any $f_1, f_2 : \mathbb{R}_+ \times \bar{\mathbb{R}}_+ \to \mathbb{R}_+$ that satisfy $f_1(x,T) \leq f_2(x,T)$, we have that $Jf_1(x,T) \leq Jf_2(x,T)$ for all $(x,T) \in \mathbb{R}_+ \times \bar{\mathbb{R}}_+$.*

As we shall see next, the operator J preserves convexity (with respect to the first variable).

Lemma 2.4. *If $f : \mathbb{R}_+ \times \bar{\mathbb{R}}_+ \to \mathbb{R}_+$ is a convex function in its first variable, then so is $Jf : \mathbb{R}_+ \times \bar{\mathbb{R}}_+ \to \mathbb{R}_+$.*

Proof. Note that Jf can be written as

(2.14)
$$Jf(x,T) = \sup_{\tau\in\mathcal{S}_{0,T}} \mathbb{E} \left\{ \int_0^{\tau} e^{-(r+\lambda)t}\lambda \cdot Pf(xH_t, T-t)dt + e^{-(r+\lambda)\tau}(K - xH_\tau)^+ \right\}.$$

Since $f(\cdot, T-t)$ is convex, so is $Pf(\cdot, T-t)$. As a result the integral with respect to time in (2.14) is also convex in x. On the other hand, note that $(K-xH_\tau)^+$ is also a convex function of x. Taking the expectation does not change the convexity with respect to x. Since the upper envelope (supremum) of convex functions is convex, the result follows. □

Remark 2.2. Since $x = 0$ is an absorbing boundary for the process S^0, for any $f : \mathbb{R}_+ \times \bar{\mathbb{R}}_+ \to \mathbb{R}_+$,

$$\begin{aligned} Jf(0,T) &= \sup_{t \in \{0,T\}} \left\{ \int_0^t e^{-(r+\lambda)s} \lambda f(0, T-s) ds + e^{-(\lambda+r)t} K \right\} \\ &= \max \left\{ K, \int_0^T e^{-(r+\lambda)s} \lambda f(0, T-s) ds + e^{-(\lambda+r)T} K \right\}, \quad T \geq 0. \end{aligned} \tag{2.15}$$

If we further assume that $f \leq K$, then $Jf(0,T) = K$, $T \geq 0$.

Lemma 2.5. *Let us assume that $f : \mathbb{R}_+ \times \bar{\mathbb{R}}_+ \to \mathbb{R}_+$ is convex in its first variable and $\|f\|_\infty \leq K$. Then $x \to Jf(x,t)$ satisfies*

$$|Jf(x,T) - Jf(y,T)| \leq |x - y|, \quad (x,y) \in \mathbb{R}_+ \times \bar{\mathbb{R}}_+, \tag{2.16}$$

and all $T \geq 0$.

Proof. First note that a positive convex function that is bounded from above has to be nonincreasing. Therefore f is nonincreasing. As a result of Lemma 2.1, $x \to Jf(x,t)$ is nonincreasing. This function is convex (by Lemma 2.4), and it satisfies

$$Jf(x,T) \geq (K-x)^+, \quad Jf(0,T) = K. \tag{2.17}$$

Consequently, the left and right derivatives of Jf satisfy

$$-1 \leq D_-^x Jf(x,T) \leq D_+^x Jf(x,T) \leq 0, \quad x > 0,\ T \geq 0. \tag{2.18}$$

Now, the result follows since the derivatives are bounded by 1 (also see Theorem 24.7 (on page 237) in [17]). □

Remark 2.3. Let $T_0 \in (0, \infty)$ and denote

$$F(x,T) = \sup_{\tau \in \mathcal{S}_{0,T}} \mathbb{E}\left\{ e^{-(r+\lambda)\tau} (K - xH_\tau)^+ \right\}, \quad x \in R_+,\ T \in [0, T_0]. \tag{2.19}$$

Then for $S \leq T \leq T_0$

$$F(x,T) - F(x,S) \leq C \cdot |T-S|^{1/2} \tag{2.20}$$

for all $x \in \mathbb{R}_+$ and for some C that depends only on T_0. See, e.g., equation (2.4) in [14].

The next lemma, which is very crucial for our proof of the smoothness of the American option price for jump diffusions, shows that the increase in the Hölder seminorm that the operator J causes can be controlled.

Lemma 2.6. *Let us assume that for some $L \in (0, \infty)$*

$$|f(x,T) - f(x,S)| \leq L|T-S|^{1/2}, \quad (T,S) \in [S_0, T_0] \times [S_0, T_0], \tag{2.21}$$

for all $x \in \mathbb{R}_+$ and for $0 \leq S_0 < T_0 < \infty$. Then

$$|Jf(x,T) - Jf(x,S)| \leq (a\,L + C)\,|T-S|^{1/2}, \quad (T,S) \in [S_0, T_0] \times [S_0, T_0], \tag{2.22}$$

for some $a \in (0,1)$ whenever

$$|T - S| < \left(\frac{r}{r+\lambda} \frac{L}{\lambda K} \right)^2. \tag{2.23}$$

Here, $C \in (0, \infty)$ is as in Remark 2.3.

Proof. Without loss of generality we will assume that $T > S$. Then we can write

$$
\begin{aligned}
&Jf(x,T) - Jf(x,S)\\
&\leq \sup_{\tau \in \mathcal{S}_{0,T}} \Bigg[\mathbb{E}\left\{\int_0^\tau e^{-(r+\lambda)t}\lambda\, Pf(xH_t, T-t)dt + e^{-(r+\lambda)\tau}(K - xH_\tau)^+\right\}\\
&- \mathbb{E}\bigg\{ \int_0^{\tau\wedge S} e^{-(r+\lambda)t}\lambda\, Pf(xH_t, S-t)dt + e^{-(r+\lambda)(\tau\wedge S)}(K - xH_{\tau\wedge S})^+\bigg\}\Bigg]\\
&= \sup_{\tau \in \mathcal{S}_{0,T}} \Bigg[\mathbb{E}\bigg\{ \int_0^\tau e^{-(r+\lambda)t}\lambda\, \left(Pf(xH_t, T-t) - Pf(xH_t, S-t)\right) dt\\
&+ 1_{\{S<\tau\}} \bigg[\int_S^\tau e^{-(r+\lambda)t}\lambda\, Pf(xH_t, S-t)dt\\
&\qquad + \left(e^{-(r+\lambda)\tau}(K - xH_\tau)^+ - e^{-(r+\lambda)S}(K - xH_S)^+\right)\bigg]\bigg\}\Bigg]\\
&\leq \frac{\lambda}{r+\lambda}L\,(T-S)^{1/2} + \frac{\lambda}{r+\lambda}K\,\left(e^{-(r+\lambda)S} - e^{-(r+\lambda)T}\right)\\
&\qquad + \sup_{\tau\in\mathcal{S}_{S,T}} \mathbb{E}\left\{e^{-(r+\lambda)\tau}(K - xH_\tau)^+\right\} - \mathbb{E}\left\{e^{-(r+\lambda)S}(K - xH_S)^+\right\}\\
&\leq \frac{\lambda}{r+\lambda}L\,(T-S)^{1/2} + \lambda K\,(T-S) + e^{-(r+\lambda)S}\left(F(H_S, T-S) - F(H_S, 0)\right)\\
&\leq \left(\frac{\lambda}{r+\lambda}L + C\right)\,(T-S)^{1/2} + \lambda K\,(T-S),
\end{aligned}
\tag{2.24}
$$

in which F is given by (2.19). To derive the second inequality in (2.24), we use the fact that

$$
\begin{aligned}
|Pf(xH_t, T-t) - Pf(xH_t, S-t)| \leq \int_{\mathbb{R}_+} \nu(dz)|f(xzH_t, T-t)\\
- f(xzH_t, S-t)| \leq L\,|T-S|^{1/2},
\end{aligned}
\tag{2.25}
$$

which follows from the assumption in (2.21), and that

$$
\begin{aligned}
&\mathbb{E}\left\{1_{\{S<\tau\}} \int_0^{\tau\wedge S} e^{-(r+\lambda)t}\lambda\, Pf(xH_t, S-t)dt\right\}\\
&\qquad \leq \lambda K \mathbb{E}\left\{\int_S^T e^{-(r+\lambda)t}dt\right\} \leq \frac{\lambda K}{\lambda + K}\left(e^{-(r+\lambda)S} - e^{-(r+\lambda)T}\right).
\end{aligned}
\tag{2.26}
$$

To derive the third inequality in (2.24), we use

$$
e^{-(r+\lambda)S} - e^{-(r+\lambda)T} \leq e^{-(r+\lambda)S}(r+\lambda)(T-S) \leq (r+\lambda)(T-S).
\tag{2.27}
$$

The last inequality in (2.24) follows from (2.20). Equation (2.22) follows from (2.24) whenever T and S satisfy (2.23). □

Let us define the continuation region and its sections by

$$\mathcal{C}^{Jf} := \{(T,x) \in (0,\infty)^2 : Jf(x,T) > (K-x)^+\}, \tag{2.28}$$

$$\mathcal{C}^{Jf}_T := \{x \in (0,\infty) : Jf(T,x) > (K-x)^+\},$$

$T > 0$, respectively.

Lemma 2.7. *Suppose that $f : \mathbb{R}_+ \times \bar{\mathbb{R}}_+ \to \bar{\mathbb{R}}_+$ is such that $x \to f(x,T)$ is a positive convex function, $T \to f(x,T)$ is nondecreasing, and $\|f\|_\infty \leq K$. Then for every $T > 0$ there exists $c^{Jf}(T) \in (0,K)$ such that $\mathcal{C}^{Jf}_T = (c^{Jf}(T),\infty)$. Moreover, $T \to c^{Jf}(T)$ is nonincreasing.*

Proof. Let us first show that if $x \geq K$, then $x \in \mathcal{C}^{Jf}_T$ for all $T \geq 0$. Let $\tau_\varepsilon := \inf\{0 \leq t \leq T : S^0_t \leq K - \varepsilon\}$. Since $\mathbb{P}\{0 < \tau_\varepsilon < T\} > 0$ for $x \geq K$, for all $T > 0$, we have that

$$\mathbb{E}^x \left\{ \int_0^{\tau_\varepsilon} e^{-(r+\lambda)t} \lambda \, Pf(S^0_t, T-t)dt + e^{-(r+\lambda)\tau_\varepsilon}(K - S^0_{\tau_\varepsilon})^+ \right\} > 0, \tag{2.29}$$

which implies that $x \in \mathcal{C}^{Jf}_T$. On the other hand, it is clear that

$$(K-x)^+ \leq Jf(x,T) \leq Jf(x,\infty), \quad (x,T) \in \mathbb{R}_+ \times \bar{\mathbb{R}}_+. \tag{2.30}$$

Thanks to Lemma 2.6 of [3], there exist $l^f \in (0,K)$ such that

$$Jf(x,\infty) = (K-x)^+, \ x \in [0,l^f]; \quad Jf(x,\infty) > (K-x)^+, \ x \in (l^f,\infty). \tag{2.31}$$

Since $x \to Jf(x,\infty)$ and $x \to Jf(x,T)$, $T \geq 0$, are convex functions (from Lemma 2.2 in [3] and Lemma 2.4, respectively), (2.29), (2.30), and (2.31) imply that there exists a point $c^{Jf}(T) \in (l^f, K)$ such that

$$Jf(x) = (K-x)^+, \ x \in [0, c^{Jf}(T)]; \quad Jf(x,T) > (K-x)^+, \ x \in (c^{Jf}(T),\infty), \tag{2.32}$$

for $T > 0$. This proves the first statement of the lemma. The fact that $T \to c(T)$ is nonincreasing follows from the fact that $T \to Jf(x,T)$ is nondecreasing. □

In the following lemma we will argue that if f has certain regularity properties, then Jf is the classical solution of a parabolic free boundary equation.

Lemma 2.8. *Let us assume that $f : \mathbb{R}_+ \times \bar{\mathbb{R}}_+ \to \mathbb{R}_+$ is convex in its first variable, $\|f\|_\infty \leq K$, and $T \to f(x,T)$ is nonincreasing. Moreover, we will assume that f satisfies*

$$|f(x,T) - f(x,S)| \leq A\,|T-S|^{1/2} \quad \textit{whenever} \quad |T-S| < B \tag{2.33}$$

for all $x \in \mathbb{R}_+$, where A, B are strictly positive constants that do not depend on x. Then the function $Jf : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ is the unique bounded solution (in the classical sense) of

$$\mathcal{A}u(x,T) - (r+\lambda)\cdot u(x,T) + \lambda \cdot (Pf)(x,T) - \frac{\partial}{\partial T}u(x,T) = 0, \quad x > c^{Jf}(T), \tag{2.34}$$

$$u(x,T) = (K-x), \quad x \leq c^{Jf}(T), \tag{2.35}$$

in which $\mathcal{A}$ is as in (2.5) *and c^{Jf} is as in Lemma* 2.7. *Moreover,*

$$\mathcal{A}Jf(x,T)-(r+\lambda)\cdot Jf(x,T)+\lambda\cdot(Pf)(x,T)-\frac{\partial}{\partial T}Jf(x,T)\leq 0, \quad x<c^{Jf}(T). \tag{2.36}$$

Proof. The proof is motivated by Theorem 2.7.7 of [10]. Equation (2.35) is clearly satisfied by Jf. In what follows, we will first show that Jf satisfies (2.34). Let us take a point in $(t,T)\in\mathcal{C}^{Jf}$ and consider a bounded rectangle $R=(t_1,t_2)\times(x_1,x_2)$ containing this point. We will let

$$t_2-t_1<B\wedge\left(\frac{rA}{(r+\lambda)\lambda K}\right)^2. \tag{2.37}$$

Let $\partial_0 R$ be the parabolic boundary of R and consider the parabolic partial differential equation

$$\begin{aligned}&\mathcal{A}u(x,T)-(r+\lambda)\cdot u(x,T)+\lambda\cdot(Pf)(x,T)-\frac{\partial}{\partial T}u(x,T)=0 \quad \text{in}\, R,\\ &u(x,T)=Jf(x,T) \quad \text{on}\ \partial_0 R.\end{aligned} \tag{2.38}$$

As a result of Lemmas 2.5 and 2.6, Jf satisfies the uniform Lipschitz and Hölder continuity conditions, which implies that Jf is continuous. On the other hand, for any $(T,x)\in R$

$$\begin{aligned}|Pf(x,T)-Pf(y,S)|&\leq|Pf(x,T)-Pf(x,S)|+|Pf(x,S)-Pf(y,S)|\\ &\leq\int_{\mathbb{R}_+}\nu(dz)\left(|f(xz,T)-f(xz,S)|+|f(xz,S)-f(yz,S)|\right)\\ &\leq A\,|T-S|^{1/2}+\xi\,|x-y|.\end{aligned} \tag{2.39}$$

Now, Theorem 5.2 in [8] implies that (2.38) has a unique classical solution. We will show that this unique solution coincides with Jf using the optional sampling theorem. Let us introduce the stopping time

$$\tau:=\inf\{\theta\in[0,t_0-t_1):(t_0-\theta,x_0H_\theta\in\partial_0 R\}\wedge(t_0-t_1), \tag{2.40}$$

which is the first time S^0 hits the parabolic boundary when S^0 starts from (x_0,t_0). Let us also define the process $N_\theta:=e^{-(r+\lambda)\theta}u(x_0H_\theta,t_0-\theta)+\int_0^\theta e^{-(r+\lambda)t}\lambda\cdot Pf(S_t^0,t_0-t)dt$, $\theta\in[0,t_0-t_1]$. From the classical Itô formula it follows that the stopped process $N_{\theta\wedge\tau}$ is a bounded martingale. As a result

$$\begin{aligned}u(x_0,t_0)&=N_0=\mathbb{E}^x\{N_\tau\}\\ &=\mathbb{E}\left\{e^{-(r+\lambda)\tau}Jf(xH_\tau,t_0-\tau))+\int_0^\tau e^{-(r+\lambda)t}\lambda\cdot Pf(S_t^0,t_0-t)dt\right\}.\end{aligned} \tag{2.41}$$

Clearly $\tau\leq\tau_x$. Since the stopped process $(e^{-(r+\lambda)(t\wedge\tau_x)}Jf(S^0_{t\wedge\tau_x},t_0-t\wedge\tau_x)+\int_0^{t\wedge\tau_x}e^{-(r+\lambda)t}\lambda\cdot Pf(S_s^0,t_0-s)ds)_{t\geq 0}$ is a bounded martingale, another application of the optional sampling theorem yields

(2.42)
$$\mathbb{E}\left\{e^{-(r+\lambda)\tau}Jf(x_0H_\tau,t_0-\tau)+\int_0^\tau e^{-(r+\lambda)t}\lambda\cdot Pf(S_t^0,t_0-t)dt\right\}=Jf(x_0,t_0).$$

Combining (2.41) and (2.42), we see that (2.34) is satisfied in the classical sense since the choice of $(x_0,t_0)\in\mathcal{C}^{Jf}$ is arbitrary.

We still need to show uniqueness among bounded functions. Fix $x>c^{Jf}(T)$. Let u be a bounded function satisfying (2.34) and (2.35). Let us define $M_t:=e^{-(r+\lambda)t}u(xH_t,T-t)+\int_0^t e^{-(r+\lambda)t}\lambda\cdot Pf(S_s^0,T-s)ds$. Using the classical Itô formula it can be seen that $M_{t\wedge\tau_x}$ is a bounded martingale. Since τ_x is optimal (see (2.11)), by the optional sampling theorem, we have

(2.43)
$$\begin{aligned}u(x,T)&=M_0=\mathbb{E}^x\{M_{\tau_x}\}\\&=\mathbb{E}\left\{e^{-(r+\lambda)\tau}u(xH_{\tau_x},T-\tau_x)+\int_0^{\tau_x}e^{-(r+\lambda)t}\lambda\cdot Pf(S_s^0,T-s)ds\right\}\\&=\mathbb{E}\left\{e^{-(r+\lambda)\tau}(K-xH_{\tau_x})^++\int_0^{\tau_x}e^{-(r+\lambda)t}\lambda\cdot Pf(S_s^0,T-s)ds\right\}=Jf(x,T).\end{aligned}$$

Next, we will prove (2.36). To this end, let $x<c^{Jf}(t)$. Let U be a closed interval centered at x such that $U\subset(0,c^{Jf}(T))$. Let $\tau_U=\{t\geq 0: xH_t\notin U\}$. Since $(e^{-(r+\lambda)t}Jf(S_t^0,T-t)+\int_0^t e^{-(r+\lambda)s}\lambda\cdot Pf(S_s^0,T-s)ds)_{t\geq 0}$ is a supermartingale, we can write

$$\begin{aligned}\mathbb{E}\Bigg[&e^{-(r+\lambda)(\tau_U\wedge t)}Jf(xH_{\tau_U\wedge t},T-\tau_U\wedge t))\\&+\int_0^{\tau_U\wedge t}e^{-(r+\lambda)u}\lambda Pf(xH_u,T-u)du\Bigg]\leq Jf(x,T)\end{aligned}\tag{2.44}$$

for all $t\geq 0$. Since $Jf(x,t)=K-x$ when $(T,x)\in\mathbb{R}_+^2-C^{Jf}$, we can apply Itô's formula to obtain that

$$\begin{aligned}\lim_{t\to 0}\mathbb{E}\Bigg[\frac{1}{t}\int_0^{\tau_U\wedge t}&e^{-(r+\lambda)u}\left(\left(\mathcal{A}-(r+\lambda)\cdot-\frac{\partial}{\partial T}\right)\right.\\&\left.Jf(xH_u,t-u)+\lambda Pf(xH_u,T-u)\right)du\Bigg]\leq 0.\end{aligned}\tag{2.45}$$

Now, (2.36) follows thanks to the dominated convergence theorem, which allows us to exchange the limit and the expectation. We can apply the dominated convergence theorem thanks to the fact that U is a compact domain. □

LEMMA 2.9. *For a given $T>0$, let $x\to f(x,T)$ be a convex and nonincreasing function. Then the convex function $x\to Jf(x,T)$ is of class C^1 at $x=c(T)$, i.e.,*

$$\left.\frac{\partial}{\partial x}Jf(x,T)\right|_{x=c(T)}=-1.\tag{2.46}$$

Proof. The proof is similar to the proof of Lemma 7.8 on page 74 of [10], but we will provide it here for the sake of completeness. If we let $x = c(T)$, then

$$\begin{aligned}
&(2.47)\\
&Jf(x+\varepsilon, T)\\
&\quad = \mathbb{E}\left\{\int_0^{\tau_{x+\varepsilon}} e^{-(r+\lambda)t}\lambda \cdot Pf((x+\varepsilon)H_t, T-t)dt + e^{-(r+\lambda)\tau_{x+\varepsilon}}(K-(x+\varepsilon)H_{\tau_{x+\varepsilon}})^+\right\}\\
&\quad = \mathbb{E}\left\{\int_0^{\tau_{x+\varepsilon}} e^{-(r+\lambda)t}\lambda \cdot Pf(xH_t, T-t)dt + e^{-(r+\lambda\tau_{x+\varepsilon})}(K-xH_{\tau_{x+\varepsilon}})^+\right\}\\
&\quad + \mathbb{E}\left\{\int_0^{\tau_{x+\varepsilon}} e^{-(r+\lambda)t}\lambda \cdot \left[Pf((x+\varepsilon)H_t, T-t) - Pf(xH_t, T-t)\right]dt\right\}\\
&\quad + \mathbb{E}\left\{e^{-(r+\lambda)\tau_{x+\varepsilon}}\left[(K-(x+\varepsilon)H_{\tau_{x+\varepsilon}})^+ - (K-xH_{\tau_{x+\varepsilon}})^+\right]\right\}\\
&\quad \leq Jf(x,T) + \mathbb{E}\left\{1_{\{\tau_{x+\varepsilon}<T\}}e^{-(r+\lambda)\tau_{x+\varepsilon}}\left[(K-(x+\varepsilon)H_{\tau_{x+\varepsilon}}) - (K-xH_{\tau_{x+\varepsilon}})\right]\right\}\\
&\quad + \mathbb{E}\left\{1_{\{\tau_{x+\varepsilon}=T\}}e^{-(r+\lambda)\tau_{x+\varepsilon}}\left[(K-(x+\varepsilon)H_{\tau_{x+\varepsilon}})^+ - (K-xH_{\tau_{x+\varepsilon}})^+\right]\right\}\\
&\quad \leq Jf(x,T) - \varepsilon\mathbb{E}^x\left\{1_{\{\tau_{x+\varepsilon}<T\}}e^{-(r+\lambda)\tau_{x+\varepsilon}}H_{\tau_{x+\varepsilon}}\right\}\\
&\quad = Jf(x,T) - \varepsilon\mathbb{E}^x\left\{e^{-(r+\lambda)\tau_{x+\varepsilon}}H_{\tau_{x+\varepsilon}}\right\} + \varepsilon\mathbb{E}^x\left\{1_{\{\tau_{x+\varepsilon}=T\}}e^{-(r+\lambda)T}H_T\right\}.
\end{aligned}$$

The first inequality follows since $\tau_{x+\varepsilon}$ is not optimal when S^0 starts at x and $x \to Pf(x,T)$ is a decreasing function for any $T \geq 0$. From (2.47) it follows that

$$D_+^x Jf(x+, T) \leq -1, \tag{2.48}$$

since $e^{-(r+\lambda)t}H_t$ is a uniformly integrable martingale and $\tau_{x+\varepsilon} \downarrow 0$. Convexity of $Jf(t,x)$ (Lemma 2.4) implies that

$$-1 = D_-^x Jf(x-, t) \leq D_+^x Jf(x+, t) \leq -1, \tag{2.49}$$

which yields the desired result. □

3. A sequence of functions approximating *V*. Let us define a sequence of functions by the following iteration:

$$\begin{aligned}
&(3.1)\\
&v_0(x,T) = (K-x)^+, \quad v_{n+1}(x,T) = Jv_n(x,T),\ n \geq 0, \quad \text{for all } (x,T) \in \mathbb{R}_+ \times \mathbb{R}_+.
\end{aligned}$$

We extend these functions onto $\mathbb{R}_+ \times \bar{\mathbb{R}}_+$ by letting

$$v_n(x,\infty) = \lim_{T\to\infty} v_n(x,T). \tag{3.2}$$

This sequence of functions is a bounded sequence, as the next corollary shows.

Corollary 3.1. *For all $n \geq 0$,*

$$(K-x)^+ \leq v_n(x,T) \leq \left(1+\frac{\lambda}{r}\right)K, \quad (x,T) \in \mathbb{R}_+ \times \bar{\mathbb{R}}_+. \tag{3.3}$$

Proof. The first inequality follows since it may not be optimal to stop immediately. Let us prove the second inequality using an induction argument: Observe that $v_0(x,T) = (K-x)^+$, $(x,T) \in \mathbb{R}_+ \times \bar{\mathbb{R}}_+$, satisfies (3.3). Let us assume that (3.3) holds for n and show that it holds for $n+1$. Using (2.13), we get that

$$\|v_{n+1}\|_\infty = \|Jv_n\|_\infty \leq K + \frac{\lambda}{r+\lambda}\left(1+\frac{\lambda}{r}\right)K = \left(1+\frac{\lambda}{r}\right). \qquad \square \tag{3.4}$$

As a corollary of Lemmas 2.3 and 2.4 we can state the following corollary, whose proof can be carried out by induction.

COROLLARY 3.2. *The sequence $(v_n(x,T))_{n\geq 0}$ is increasing for all $(x,T) \in \mathbb{R}_+ \times \bar{\mathbb{R}}_+$. For each n, the function $x \to v_n(x,T)$, $x \geq 0$, is convex for all $T \in \bar{\mathbb{R}}_+$.*

Remark 3.1. Let us define

$$v_\infty(x,T) := \sup_{n\geq 0} v_n(x,T), \quad (x,T) \in \mathbb{R}_+ \times \bar{\mathbb{R}}_+. \tag{3.5}$$

This function is well defined as a result of (3.3) and Corollary 3.2. In fact, it is convex, because it is the upper envelope of convex functions, and it is bounded by the right-hand side of (3.3).

COROLLARY 3.3. *For each $n \geq 0$ and $t \in \mathbb{R}_+$, $x \to v_n(x,T)$ is a decreasing function on $[0,\infty)$. Moreover, $T \to v_n(x,T)$ is nondecreasing. The same statements hold for $x \to v_\infty(x,T)$ and $T \to v_\infty(x,T)$, respectively.*

Proof. The behavior with respect to the first variable is a result of Corollary 3.2 and Remark 3.1 since any positive convex function that is bounded from above is decreasing. For each n, the fact that $T \to v_n(x,T)$ is nondecreasing is a corollary of Lemma 2.1. On the other hand, for any $T \geq S \geq 0$, we have that $v_\infty(x,T) = \sup_n v_n(x,T) \geq \sup_n v_n(x,S) = v_\infty(x,S)$. $\square$

Next, we will sharpen the upper bound in Corollary 3.1. This improvement has some implications for the continuity of $x \to v_n(x,T)$, $n \geq 1$, and $x \to v_\infty(x,T)$ at $x = 0$.

Remark 3.2. The upper bound in (3.1) can be sharpened using Corollary 3.3 and Remark 2.2. Indeed, we have

$$\begin{aligned} &(K-x)^+ \leq v_n(x,T) < K, \quad \text{for each } n, \quad \text{and} \\ &(K-x)^+ \leq v_\infty(x,T) < K, \quad (x,T) \in (0,\infty)^2. \end{aligned} \tag{3.6}$$

It follows from this observation that for every $T \in \bar{\mathbb{R}}_+$, $x \to v_n(x,T)$, for every n, and $x \to v_\infty(x,T)$ are continuous at $x = 0$ since $v_n(0,T) = v_\infty(0,T) = K$ and these functions are convex. (Note that convexity already guarantees continuity for $x > 0$.)

LEMMA 3.1. *The function v_∞ is the smallest fixed point of the operator J.*

Proof.

$$\begin{aligned} &v_\infty(x,T-t) \\ &\quad = \sup_{n\geq 1} v_n(x,T-t) \\ &\quad = \sup_{n\geq 1} \sup_{\tau\in\mathcal{S}_{0,T}} \mathbb{E}^x \left\{ \int_0^\tau e^{-(r+\lambda)t} \lambda \cdot P v_n(S_t^0, T-t)dt + e^{-(r+\lambda)\tau}(K-S_\tau^0)^+ \right\} \end{aligned} \tag{3.7}$$

$$= \sup_{\tau\in\mathcal{S}_{0,T}} \sup_{n\geq 1} \mathbb{E}^x \left\{ \int_0^\tau e^{-(r+\lambda)t}\lambda \cdot Pv_n(S_t^0, T-t)dt + e^{-(r+\lambda)\tau}(K-S_\tau^0)^+ \right\}$$

$$= \sup_{\tau\in\mathcal{S}_{0,T}} \mathbb{E}^x \left\{ \int_0^\tau e^{-(r+\lambda)t}\lambda \cdot P(\sup_{n\geq 1} v_n)(S_t^0, T-t)dt + e^{-(r+\lambda)\tau}(K-S_\tau^0)^+ \right\}$$

$$= Jv_\infty(x, T-t),$$

in which the fourth equality follows by applying the monotone convergence theorem three times. Let $w : \mathbb{R}_+ \times \bar{\mathbb{R}}_+ \to \mathbb{R}_+$ be another fixed point of the operator J. We will argue by induction that $w \geq v_\infty$. For $(x,t) \in \mathbb{R}_+ \times \bar{\mathbb{R}}_+$, $w(x,T-t) = Jw(x,T-t)$, which implies that $w(x,T-t) = Jw(x,T-t) \geq (K-x)^+ = v_0(\cdot)$. If we assume that $w(x,T-t) \geq v_n(x,T-t)$, then $w(x,T-t) = Jw(x,T-t) \geq Jv_n(x,T-t) = v_{n+1}(x,T-t)$. Consequently $w(x,T-t) \geq v_n(x,T-t)$ for all $n \geq 0$. As a result $w(x,T-t) \geq \sup_{n\geq 0} v_n(x,T-t) = v_\infty(x,T-t)$. □

LEMMA 3.2. *The sequence $\{v_n(\cdot,\cdot)\}_{n\geq 0}$ converges uniformly to v_∞. In fact, the rate of convergence is exponential:*

$$v_n(x,T) \leq v_\infty(x,T) \leq v_n(x,T) + \left(\frac{\lambda}{\lambda+r}\right)^n K, \quad (x,T) \in \mathbb{R}_+ \times \bar{\mathbb{R}}_+. \tag{3.8}$$

Proof. The first inequality follows from the definition of v_∞. The second inequality can be proved by induction. The inequality holds when we set $n = 0$ by Remark 3.2. Assume that the inequality holds for $n > 0$. Then

(3.9)

$$v_\infty(x,T) = \sup_{\tau\in\mathcal{S}_{0,T}} \mathbb{E}^x \left\{ \int_0^\tau e^{-(r+\lambda)t}\lambda \cdot Pv_\infty(S_t^0, T-t)dt + e^{-(r+\lambda)\tau}(K-S_\tau^0)^+ \right\}$$

$$\leq \sup_{\tau\in\mathcal{S}_{0,T}} \mathbb{E}^x \left\{ \int_0^\tau e^{-(r+\lambda)t}\lambda \cdot Pv_n(S_t^0, T-t)dt + e^{-(r+\lambda)\tau}(K-S_\tau^0)^+ \right\}$$

$$+ \int_0^\infty dt\, e^{-(\lambda+r)t}\lambda \left(\frac{\lambda}{\lambda+r}\right)^n K$$

$$= v_{n+1}(x,T) + \left(\frac{\lambda}{\lambda+r}\right)^{n+1} K. \qquad \square$$

Remark 3.3. Note that, for a fixed $T_0 > 0$,

(3.10) $v_n(x,T) \leq v_\infty(x,T)$

$$\leq v_n(x,T) + \left(1 - e^{-(r+\lambda)T_0}\right)^n \left(\frac{\lambda}{\lambda+r}\right)^n K, \quad x \in \mathbb{R}_+,\ T \in (0, T_0).$$

This can be derived using an induction argument similar to the one used in the proof of Lemma 3.2. We simply replace (3.9) by

$$\begin{aligned} v_\infty(x,T) &\le \sup_{\tau\in\mathcal{S}_{0,T}} \mathbb{E}^x\left\{ \int_0^\tau e^{-(r+\lambda)t}\lambda\cdot Pv_n(S^0_t,T-t)dt + e^{-(r+\lambda)\tau}(K-S^0_\tau)^+\right\} \\ &\quad + \int_0^{T_0} dt\, e^{-(\lambda+r)t}\left(1-e^{-(r+\lambda)T_0}\right)^n \lambda\left(\frac{\lambda}{\lambda+r}\right)^n K \\ &= v_{n+1}(x,T) + K\left(1-e^{-(r+\lambda)T_0}\right)^{n+1}\left(\frac{\lambda}{\lambda+r}\right)^{n+1}. \end{aligned} \tag{3.11}$$

Observe that one can replace K in (3.10) by $\|v_\infty - v_0\|_\infty$. Note that the convergence rate in (3.10) is fast. This will lead to a numerical scheme, whose error versus accuracy characteristics can be controlled, for pricing American options.

Remark 3.4. Let $T_0 \in (0,\infty)$. It can be shown using similar arguments to the ones used in the proof of Lemma 2.6 that

$$|v_1(x,T) - v_1(x,S)| \le \lambda\, K\, |T-S| + C\,|T-S|^{1/2}, \quad T,S\in(0,T_0], \tag{3.12}$$

for all $x \in \mathbb{R}_+$, in which $C \in (0,\infty)$ is as in Remark 2.3. In fact,

$$|v_1(x,T) - v_1(x,S)| \le L\,|T-S|^{1/2} \tag{3.13}$$

for all $x \in \mathbb{R}_+$ and for some L that depends only on T_0.

The next lemma shows that the functions v_n, $n \ge 0$, and v_∞ are locally Hölder continuous with respect to the time variable.

Lemma 3.3. *Let $T_0 \in (0,\infty)$ and $L \in (0,\infty)$ be as in Remark* 3.4 *and $C \in (0,\infty)$ be as in Remark* 2.3. *Then, for $T,S \in (0,T_0)$, we have that*

$$|v_n(x,T) - v_n(x,S)| \le \left(L + \frac{C}{1-a}\right)|T-S|^{1/2} \text{ whenever } |T-S| \le \left(\frac{r}{r+\lambda}\frac{L}{\lambda K}\right)^2 \tag{3.14}$$

for all $x \in \mathbb{R}_+$ and for all $n \ge 1$. Here, $a \in (0,1)$ is as in Lemma 2.6. *Moreover,*

$$|v_\infty(x,T) - v_\infty(x,S)| \le \left(L + \frac{C}{1-a}\right)|T-S|^{1/2} \text{ whenever } |T-S| \le \left(\frac{r}{r+\lambda}\frac{L}{\lambda K}\right)^2 \tag{3.15}$$

for all $x \in \mathbb{R}_+$.

Proof. The proof of (3.14) will be carried out using an induction argument. Observe from Remark 3.4 that (3.14) holds for $n = 1$. Let us assume that (3.14) holds for n and show that it holds for $n+1$. Using Lemma 2.6, we have that

$$|v_{n+1}(x,T) - v_{n+1}(x,S)| \le \left(a\left(L + \frac{C}{1-a}\right) + C\right)|T-S|^{1/2} \tag{3.16}$$

for $|T-S| \le \left(\frac{r}{r+\lambda}\ \frac{L+C/(1-a)}{\lambda K}\right)^2$. It is clear that the right-hand side of (3.16) is less than that of (3.14), and

$$\frac{r}{r+\lambda}\ \frac{L+C/(1-a)}{\lambda K} \ge \frac{r}{r+\lambda}\ \frac{L}{\lambda K}, \tag{3.17}$$

from which the first statement of the lemma follows. Now, let us prove (3.15). To this end observe that

$$\begin{aligned}
&|v_\infty(x,T)-v_\infty(x,S)| \\
&\leq |v_\infty(x,T)-v_n(x,T)|+|v_n(x,T)-v_n(x,S)|+|v_\infty(x,S)-v_n(x,S)| \\
&\leq 2\left(\frac{\lambda}{\lambda+r}\right)^n K+\left(L+\frac{C}{1-a}\right)|T-S|^{1/2}
\end{aligned} \tag{3.18}$$

for any $n>1$, which follows from (3.14) and Lemma 3.2. The result follows since n on the right-hand side of (3.18) is arbitrary. □

LEMMA 3.4. *For $n \geq 0$, $|v_n(x,T)-v_n(y,T)| \leq |x-y|$, and $|v_\infty(x,T)-v_\infty(y,T)| \leq |x-y|$, $(x,y) \in \mathbb{R}_+ \times \bar{\mathbb{R}}_+$, for all $T \geq 0$.*

Proof. It follows from Remark 3.2 that $\|v_n\|_\infty \leq K$, for all $n \geq 0$, and $\|v_\infty\|_\infty \leq K$. Moreover, for each $n \geq 0$, $v_n(\cdot,T)$ is convex (for all $T \in \bar{\mathbb{R}}_+$) as a result of Corollary 3.2. On the other hand, it was pointed out in Remark 3.1 that $v_\infty(\cdot,T)$ is convex for all $T \in \mathbb{R}_+$. Since

$$v_{n+1}(x,T)=Jv_n(x,T) \quad \text{and} \quad v_\infty(x,T)=Jv_\infty(x,T), \tag{3.19}$$

the statement of the lemma follows from Lemma 2.5. □

LEMMA 3.5. *For all $T \geq 0$ and $n \geq 0$, $\mathcal{C}_T^{v_{n+1}}=(c^{v_{n+1}}(T),\infty)$ for some $c^{v_{n+1}}(T) \in (0,K)$ and $\mathcal{C}_T^{v_\infty}=(c^{v_\infty}(T),\infty)$ for some $c^{v_\infty} \in (0,K)$. The function v_{n+1} is the unique bounded solution (in the classical sense) of*

$$\begin{aligned}
&\mathcal{A}v_{n+1}(x,T)-(r+\lambda)\cdot v_{n+1}(x,T)+\lambda\cdot(Pv_n)(x,T) \\
&\quad -\frac{\partial}{\partial T}v_{n+1}(x,T)=0, \quad x>c^{v_{n+1}}(T), \\
&v_{n+1}(x,T)=(K-x), \quad x \leq c^{v_{n+1}}(T),
\end{aligned} \tag{3.20}$$

and it satisfies

$$\left.\frac{\partial}{\partial x}v_{n+1}(x,T)\right|_{x=c^{v_{n+1}}(T)}=-1, \quad T>0. \tag{3.21}$$

Moreover, v_∞ is the unique bounded solution (in the classical sense) of

$$\begin{aligned}
&\mathcal{A}v_\infty(x,T)-(r+\lambda)\cdot v_\infty(x,T)+\lambda\cdot(Pv_\infty)(x,T)-\frac{\partial}{\partial T}v_\infty(x,T)=0, \quad x>c^{v_\infty}(T), \\
&v_\infty(x,T)=(K-x), \quad x \leq c^{v_\infty}(T),
\end{aligned} \tag{3.22}$$

and it satisfies

$$\left.\frac{\partial}{\partial x}v_\infty(x,T)\right|_{x=c^{v_\infty}(T)}=-1, \quad T>0. \tag{3.23}$$

On the other hand,

$$\mathcal{A}v_\infty(x,T)-(r+\lambda)\cdot v_\infty(x,T)+\lambda\cdot(Pv_\infty)(x,T)-\frac{\partial}{\partial T}v_\infty(x,T) \leq 0, \quad x<c^{v_\infty}(T). \tag{3.24}$$

Proof. The fact that $\mathcal{C}^{v_{n+1}} = (c^{v_{n+1}}, \infty)$ and $C^{v_\infty} = (c^{v_\infty}, \infty)$ for some $c^{v_{n+1}} \in (0, K)$ and $c^{v_\infty} \in (0, K)$ follows from Lemma 2.7 since the assumptions in that lemma hold thanks to Corollaries 3.2 and 3.3; Remarks 3.1 and 3.2; and Lemma 3.1.

The partial differential equations (3.20), (3.22) and the inequality in (3.24) are satisfied as a corollary of Lemma 2.8; Corollaries 3.2 and 3.3, Remarks 3.1 and 3.2; and Lemmas 3.1 and 3.3.

Observe that since v_n is convex (Corollary 3.2) and nonincreasing (Corollary 3.3) with respect to its first variable, v_{n+1} $(= Jv_n)$ satisfies the smooth fit condition in (3.21) as a result of Lemma 2.9. The smooth fit condition in (3.23) holds for v_∞ as a result of Lemma 2.9 since v_∞ $(= Jv_\infty)$ (Lemma 3.1) and $x \to v_\infty(x, T)$ is nonincreasing and convex. □

The next lemma will be used to verify the fact that $V = v_\infty$. The classical Itô rule cannot be applied to the process $t \to v_\infty(S_t, T-t)$ since the function v_∞ may fail to be $C^{2,1}$ at $T \to c^{v_\infty}(T)$. As a result, the semimartingale decomposition of the process $t \to v_\infty(S_t, T-t)$ may contain an extra term due to the local time of the process S at the free boundary.

LEMMA 3.6. *Let $X = \{X_t; t \geq 0\}$ be a semimartingale and $b : \mathbb{R}_+ \to \mathbb{R}$ be a continuous function of bounded variation. Let $F : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ be a continuous function that is $C^{2,1}$ on $\bar{C}$ and $\bar{D}$ (it may not necessarily be $C^{1,2}$ across the boundary curve b), in which*

$$C \triangleq \{(x,t) \in \mathbb{R} \times \mathbb{R}_+ : x < b(t)\}, \quad D \triangleq \{(x,t) \in \mathbb{R} \times \mathbb{R}_+ : x > b(t)\}.$$

That is, there exist two functions $F^1, F^2 : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$, that are $C^{2,1}$ on $\mathbb{R} \times \mathbb{R}_+$, and $F(x,t) = F^1(x,t)$ when $(x,t) \in C$ and $F(x,t) = F^2(x,t)$ when $(x,t) \in D$. Moreover, $F^1(b(t),t) = F^2(b(t),t)$. Then the following generalization of Itô's formula holds:

(3.25)
$$\begin{aligned}
F(X_t, t) = F(X_0, 0) &+ \int_0^t \frac{1}{2}\left[F_t(X_{s-}+, s) + F_t(X_{s-}-, s)\right] ds \\
&+ \frac{1}{2}\int_0^t \left[F_x(X_{s-}+, s) + F_x(X_{s-}-, s)\right] dX_s \\
&+ \frac{1}{2}\int_0^t 1_{\{X_{s-} \neq b(s)\}} F_{xx}(X_{s-}, s) d\left\langle X, X\right\rangle_s^c \\
&+ \sum_{0<s\leq t} \left\{F(X_s, s) - F(X_{s-}, s) - \frac{1}{2}\Delta X_s \left[F_x(X_{s-}-, s) + F_x(X_{s-}+, s)\right]\right\} \\
&+ \frac{1}{2}\int_0^t \left[F_x(X_{s-}+, s) - F_x(X_{s-}-, s)\right] 1_{\{X_{s-}=b(s)\}} dL_t^b,
\end{aligned}$$

where L_t^b is the local time of the semimartingale $X_t - b(t)$ at zero (see the definition on page 216 *in* [16]*).*

Lemma 3.6 was stated in Theorem 2.1 of [12] for continuous semimartingales. The generalization for the case when the underlying process is not necessarily continuous is intuitively clear and just technical, but we will prove it in the appendix for the sake of completeness. We are now ready to state the main result.

THEOREM 3.1. *The value function V is the unique bounded solution (in the classical sense) of the integro-partial differential equation in* (3.22)*. Moreover, it satisfies* *the smooth fit condition at the optimal stopping boundary, i.e.,* $\frac{\partial}{\partial x}V(x,T)\big|_{x=c^{v_\infty}(T)} = -1$, $T > 0$.

Proof. The proof is a corollary of the optional sampling theorem and the generalized Itô formula given above. Let $T \in (0,\infty)$ and define

$$(3.26)\qquad \widetilde{M}_t = e^{-rt}v_\infty(S_t, T-t) \quad \text{and} \quad \widetilde{\tau}_x := T \wedge \inf\{t \in [0,T] : S_t \le c^{v_\infty}(T-t)\}.$$

It follows from (3.22) and the classical Itô lemma that $\{\widetilde{M}_{t\wedge\widetilde{\tau}_x}\}_{0\le t\le T}$ is a bounded $\mathbb{P}$-martingale. Using the optional sampling theorem, one obtains

$$(3.27)\qquad \begin{aligned} v_\infty(x,T) &= \widetilde{M}_0 = \mathbb{E}^x\left\{\widetilde{M}_{\widetilde{\tau}_x}\right\} = \mathbb{E}^x\left\{e^{-r\widetilde{\tau}_x}v_\infty(S_{\widetilde{\tau}_x}, T-\widetilde{\tau}_x)\right\} \\ &= \mathbb{E}^x\left\{e^{-r\widetilde{\tau}_x}(K - S_{\widetilde{\tau}_x})^+\right\} \le V(x,T). \end{aligned}$$

In the rest of the proof we will show that $v_\infty(x,T) \ge V(x,T)$. Since v_∞ satisfies the smooth fit principle across the free boundary, when we apply the generalized Itô formula to $v_\infty(S_t, T-t)$, the local time term drops. Thanks to (3.22) and (3.24), $v_\infty(S_t, T-t)$ is a positive $\mathbb{P}$-supermartingale. Again, using the optional sampling theorem, for any $\tau \in \widetilde{\mathcal{S}}_{0,T}$

(3.28)

$$v_\infty(x,T) = \widetilde{M}_0 \ge \mathbb{E}^x\left\{\widetilde{M}_\tau\right\} = \mathbb{E}^x\left\{e^{-r\tau}v_\infty(S_\tau, T-\tau)\right\} \ge \mathbb{E}^x\left\{e^{-r\tau}(K-S_\tau)^+\right\}.$$

As a result $v_\infty(x,T) \ge V(x,T)$. □

Remark 3.5. We have that

$$(3.29)\qquad \mathcal{C}_T^{v_\infty} = \{x \in (0,\infty) : v_\infty > (K-x)^+\} = (c^{v_\infty}(T), \infty).$$

On the other hand, $v_\infty = K - x$ for $x \le c^{v_\infty}$. Since $V = v_\infty$, by Theorem 3.1, it follows that

$$(3.30)\qquad \mathcal{C}_T^{V} = \{x \in (0,\infty) : V > (K-x)^+\} = (c^{v_\infty}(T), \infty).$$

Appendix. Proof of Lemma 3.6. As in [12] we will define $Z_t^1 = X_t \wedge b(t)$, $Z_t^2 = X_t \vee b(t)$, and observe that

$$(\text{A.1})\qquad F(X_t, t) = F^1(Z_t^1, t) + F^2(Z_t^2, t) - F(b(t), t).$$

On the other hand, applying the Meyer–Itô formula (see Theorem 70 in [16]) to the semimartingale $X_t - b(t)$, we obtain

(A.2)

$$\begin{aligned} |X_t - b(t)| &= |X_0 - b(0)| + \int_0^t \operatorname{sign}(X_{s-} - b(s))d(X_s - b(s)) \\ &\quad + 2\sum_{0<s\le t}\left[1_{\{X_{s-}>b(s)\}}(X_s - b(s))^- + 1_{\{X_s \le b(s)\}}(X_s - b(s))^+\right] + L_t^b. \end{aligned}$$

Since $Z_t^1 = \frac{1}{2}(X_t + b(t) - |X_t - b(t)|)$ and $Z_t^2 = \frac{1}{2}(X_t + b(t) + |X_t - b(t)|)$, using (A.2), we get

$$(\text{A.3})\qquad \begin{aligned} dZ_t^1 &= \frac{1}{2}\left\{(1 - \operatorname{sign}(X_{t-} - b(t)))dX_t + (1 + \operatorname{sign}(X_{t-} - b(t)))db(t) - dL_t^b\right\} \\ &\quad - \left[1_{\{X_{t-}>b(t)\}}(X_t - b(t))^- + 1_{\{X_t \le b(t)\}}(X_t - b(t))^+\right], \end{aligned}$$

$$\begin{aligned} dZ_t^2 &= \frac{1}{2}\left\{(1+\operatorname{sign}(X_{t-}-b(t)))dX_t + (1+\operatorname{sign}(X_{t-}-b(t)))db(t) - dL_t^b\right\} \\ &\quad + \left[1_{\{X_{t-}>b(t)\}}(X_t-b(t))^- + 1_{\{X_t\le b(t)\}}(X_t-b(t))^+\right]. \end{aligned} \tag{A.4}$$

It follows from the dynamics of Z^i, $i \in \{1,2\}$, that

$$d\left\langle Z^i, Z^i\right\rangle_t^c = \left(1_{\{X_{t-}<b(t)\}} + \frac{1}{4}1_{\{X_{t-}=b(t)\}}\right) d\left\langle X, X\right\rangle_t^c = 1_{\{X_{t-}<b(t)\}} d\left\langle X, X\right\rangle_t^c, \tag{A.5}$$

where the second equality follows from the occupation density formula; see, e.g., Corollary 1 on page 219 of [16]. Applying the classical Itô formula to $F^1(Z_t^1, t)$ and $F^2(Z_t^2, t)$ and using the dynamics of Z^1 and Z^2, we get

$$\begin{aligned} F^1(Z_t^1,t) &= F^1(Z_0^1,0) + \int_0^t F_t^1(Z_{s-}^1,s)ds + \int_0^t F_x^1(Z_{s-}^1,s)dZ_s^1 \\ &\quad + \frac{1}{2}\int_0^t F_{xx}^1(s, Z_{s-}^1) d\left\langle Z^1, Z^1\right\rangle_s^c \\ &\quad + \sum_{0s\le t}\left[F^1(Z_s^1,s) - F^1(Z_{s-}^1,s) - \Delta Z_s^1 F_x^1(Z_{s-}^1,s)\right] \\ &= F^1(Z_0^1,0) + \int_0^t F_t^1(Z_{s-}^1,s)ds \\ &\quad + \frac{1}{2}\int_0^t (1-\operatorname{sign}(X_{s-}-b(s)))F_x^1(Z_{s-}^1,s)dX_s \\ &\quad + \frac{1}{2}\int_0^t (1+\operatorname{sign}(X_{s-}-b(s)))F_x^1(Z_{s-}^1,s)db(s) \\ &\quad - \sum_{0<s\le t}\left[1_{\{X_{s-}>b(s)\}}(X_s-b(s))^- + 1_{\{X_s\le b(s)\}}(X_s-b(s))^+\right] F_x^1(Z_{s-}^1,s) \\ &\quad - \frac{1}{2}\int_0^t F_x^1(Z_{s-}^1,s)dL_t^b + \frac{1}{2}\int_0^t 1_{\{X_{s-}<b(s)\}} F_{xx}^1(Z_{s-}^1,s) d\left\langle X^c, X^c\right\rangle_s \\ &\quad \sum_{0<s\le t}\left[F^1(Z_s^1,s) - F(Z_{s-}^1,s) - \Delta Z_s^1 F_x^1(Z_{s-}^1,s)\right]\Big], \end{aligned} \tag{A.6}$$

$$
\begin{aligned}
F^2(Z_t^2,t) &= F^2(Z_0^2,0) + \int_0^t F_t^2(Z_{s-}^2,s)ds + \int_0^t F_x^2(Z_{s-}^2,s)dZ_s^2 \\
&\quad + \frac{1}{2}\int_0^t F_{xx}^2(s,Z_{s-}^2)d\left\langle Z^2,Z^2\right\rangle_s^c \\
&\quad + \sum_{0s\leq t}\left[F^2(Z_s^2,s) - F^2(Z_{s-}^2,s) - \Delta Z_s^2 F_x^1(Z_{s-}^2,s)\right] \\
&= F^2(Z_0^2,0) + \int_0^t F_t^2(Z_{s-}^2,s)ds \\
&\quad + \frac{1}{2}\int_0^t (1+\operatorname{sign}(X_{s-}-b(s)))F_x^2(Z_{s-}^2,s)dX_s \\
&\quad + \frac{1}{2}\int_0^t (1-\operatorname{sign}(X_{s-}-b(s)))F_x^2(Z_{s-}^2,s)db(s) \\
&\quad + \sum_{0<s\leq t}\left[1_{\{X_{s-}>b(s)\}}(X_s-b(s))^- + 1_{\{X_s\leq b(s)\}}(X_s-b(s))^+\right]F_x^2(Z_{s-}^2,s) \\
&\quad - \frac{1}{2}\int_0^t F_x^2(Z_{s-}^2,s)dL_t^b + \frac{1}{2}\int_0^t 1_{\{X_{s-}<b(s)\}}F_{xx}^2(Z_{s-}^2,s)d\left\langle X^c,X^c\right\rangle_s \\
&\quad + \sum_{0<s\leq t}\left[F^2(Z_s^2,s) - F(Z_{s-}^2,s) - \Delta Z_s^2 F_x^2(Z_{s-}^2,s)\right].
\end{aligned}
\tag{A.7}
$$

By splitting each term into its respective values on the sets $\{X_{s-}<b(s)\}$, $\{X_{s-}=b(s)\}$, and $\{X_{s-}>b(s)\}$, it can be seen that the following four equations are satisfied:

$$
F^1(Z_0^1,0) + F^2(Z_0^2,0) = F(X_0,0) + F(b(0),0),
\tag{A.8}
$$

$$
\begin{aligned}
\int_0^t F_t^1(Z_{s-},s)ds + \int_0^t F_t^2(Z_{s-}^2,s)ds &= \frac{1}{2}\int_0^t F_t(X_{s-}+,s) + F_t(X_{s-}-,s)ds \\
+ \int_0^t \Bigg[F_t(b(s)+,s)1_{\{X_{s-}<b(s)\}} &+ \frac{1}{2}(F_t(b(s)-,s) \\
+F_t(b(s)+,s))1_{\{X_{s-}=b(s)\}} &+ F_t(b(s)-,s)1_{\{X_{s-}>b(s)\}}\Bigg]ds,
\end{aligned}
\tag{A.9}
$$

$$
\begin{aligned}
&\frac{1}{2}\int_0^t (1-\operatorname{sign}(X_{s-}-b(s)))F_x^1(Z_{s-}^1,s)dX_s \\
&+ \frac{1}{2}\int_0^t (1+\operatorname{sign}(X_{s-}-b(s)))F_x^2(Z_{s-}^2,s)dX_s \\
&= \frac{1}{2}\int_0^t \left[F_x(X_{s-}+,s) + F_x(X_{s-}-,s)\right]dX_s,
\end{aligned}
\tag{A.10}
$$

$$\begin{aligned} &\frac{1}{2}\int_0^t (1+\operatorname{sign}(X_{s-}-b(s)))F_x^1(Z_{s-}^1,s)db(s) \\ &\qquad + \frac{1}{2}\int_0^t (1-\operatorname{sign}(X_{s-}-b(s)))F_x^2(Z_{s-}^2,s)db(s) \\ &\quad = \int_0^t \Bigg[F_x(b(s)+,s)1_{\{X_{s-}<b(s)\}} \\ &\qquad + \frac{1}{2}\left[F_x(b(s)+,s)+F_x(b(s)-,s)\right]1_{\{X_{s-}=b(s)\}} + F_x(b(s)-,s)1_{\{X_{s-}>b(s)\}}\Bigg]db(s). \end{aligned} \tag{A.11}$$

On the other hand, equation (3.15) of [12] still holds:

$$\begin{aligned} F(b(t),t) = F(b(0),0) + \int_0^t &\Bigg[F_t(b(s)+,s)1_{\{X_{s-}<b(s)\}} \\ &+ \frac{1}{2}\left[F_t(b(s)-,s)+F_t(b(s)+,s)1_{\{X_{s-}=b(s)\}}\right] \\ &+ F_t(b(s)-,s)1_{\{X_{s-}>b(s)\}}\Bigg]ds \\ &\int_0^t \Bigg[F_x(b(s)+,s)1_{\{X_{s-}<b(s)\}} + \frac{1}{2}\left[F_x(b(s)-,s)+F_x(b(s)+,s)1_{\{X_{s-}=b(s)\}}\right] \\ &+ F_x(b(s)-,s)1_{\{X_{s-}>b(s)\}}\Bigg]db(s), \end{aligned} \tag{A.12}$$

whose proof is carried out by using the uniqueness of finite measures on p-systems.

Let us analyze the jump terms in (A.6) and (A.7). We will denote

$$\begin{aligned} A := &-\left[1_{\{X_{s-}>b(s)\}}(X_s-b(s))^- + 1_{\{X_s\le b(s)\}}(X_s-b(s))^+\right]F_x^1(Z_{s-}^1,s) \\ &+ \left[F^1(Z_s^1,s) - F(Z_{s-}^1,s) - \Delta Z_s^1 F_x^1(Z_{s-}^1,s)\right], \end{aligned} \tag{A.13}$$

$$\begin{aligned} B := &\left[1_{\{X_{s-}>b(s)\}}(X_s-b(s))^- + 1_{\{X_s\le b(s)\}}(X_s-b(s))^+\right]F_x^2(Z_{s-}^2,s) \\ &+ \sum_{0<s\le t}\left[F^2(Z_s^2,s) - F(Z_{s-}^2,s) - \Delta Z_s^2 F_x^2(Z_{s-}^2,s)\right]. \end{aligned} \tag{A.14}$$

Depending on the whereabouts of X_{s-} and X_s with respect to the boundary curve b, A and B take four different values:

1. $X_{s-} > b(s)$ and $X_t \ge b(t)$. In this case

$$A = 0, \quad B = F^2(X_s,s) - F^2(X_{s-},s) - \Delta X_s F_x^2(X_{s-},s), \tag{A.15}$$

$$A + B = F(X_s,s) - F(X_{s-},s) - \Delta X_s F_x(X_{s-}+,s). \tag{A.16}$$

2. $X_{s-} > b(s)$ and $X_s < b(s)$. In this case

$$\begin{aligned} A &= -(b(s)-X_s)F_x^1(b(s),s) + F^1(X_s,s) \\ &\qquad - F^1(b(s),s) - (X_s-b(s))F_x^1(b(s),s) \\ &= F^1(X_s,s) - F^1(b(s),s), \end{aligned} \tag{A.17}$$

$$\begin{aligned} B &= (b(s) - X_s)F_x^2(b(s), s) + F^2(b(s), s) - F^2(X_{s-}, s) \\ &\quad - (b(s) - X_{s-})F_x^2(X_{s-}, s) \\ &= F^2(b(s), s) - F^2(X_{s-}, s) - \Delta X_s F_x^2(X_{s-}, s), \end{aligned} \tag{A.18}$$

$$A + B = F(X_s, s) - F(X_{s-}, s) - \Delta X_s F_x(X_s+, s). \tag{A.19}$$

3. $X_{s-} \leq b(s)$ and $X_s \geq b(s)$. We have that

$$\begin{aligned} A &= -(X_s - b(s))F_x^1(X_{s-}, s) + F^1(b(s), s) \\ &\quad - F^1(X_{s-}, s) - (b(s) - X_{s-})F_x^1(X_{s-}, s) \\ &= F^1(b(s), s) - F^1(X_{s-}, s) - \Delta X_s F^1(X_{s-}, s), \end{aligned} \tag{A.20}$$

$$\begin{aligned} B &= (X_s - b(s))F_x^2(b(s), s) + F^2(X_s, s) \\ &\quad - F^2(b(s), s) - (X_s - b(s))F_x^2(b(s), s) \\ &= F^2(X_s, s) - F^2(b(s), s). \end{aligned} \tag{A.21}$$

As a result

$$A + B = F(X_s, s) - F(X_{s-}, s) - \Delta X_s F_x(X_{s-}-, s). \tag{A.22}$$

4. $X_{s-} \leq b(s)$ and $X_s < b(s)$. Clearly,

$$A = F^1(X_s, s) - F^1(X_{s-}, s) - \Delta X_s F_x^1(X_{s-}, s) \quad \text{and} \quad B = 0. \tag{A.23}$$

As a result

$$A + B = F(X_s, s) - F(X_{s-}, s) - \Delta X_s F_x(X_{s-}, s). \tag{A.24}$$

Now, combining (A.1), (A.5), (A.6), (A.7), (A.8), (A.9), (A.10), (A.11), (A.12), (A.16), (A.19), (A.22), and (A.24), we obtain

$$\begin{aligned} F(X_t, t) &= F(X_0, 0) + \frac{1}{2}\int_0^t [F_t(X_{s-}+, s) + F_t(X_{s-}-, s)]\, ds \\ &\quad + \frac{1}{2}\int_0^t [F_x(X_{s-}+, s) + F_x(X_{s-}-, s)]\, dX_s \\ &\quad + \frac{1}{2}\int_0^t 1_{\{X_{s-} \leq b(s)\}} F_{xx}(s, X_{s-}) d\left\langle X, X\right\rangle_{s-}^c \\ &\quad + \sum_{0<s\leq t}\Bigl[F(X_s, s) - F(X_{s-}, s) - \Delta X_s F_x(X_{s-}-, s) 1_{\{X_{s-} \leq b(s)\}} \\ &\qquad\qquad - \Delta X_s F_x(s, X_{s-}+) 1_{\{X_{s-} > b(s)\}}\Bigr] \\ &\quad + \frac{1}{2}\int_0^t \left[F_x^2(Z_{s-}^2, s) - F^1(Z_{s-}^1, s)\right] dL_t^b. \end{aligned} \tag{A.25}$$

The last term on the right-hand side of (A.25) can be written as

$$\begin{aligned}\text{(A.26)} \qquad & \frac{1}{2}\int_0^t \left[F_x^2(Z_{s-}^2, s) - F^1(Z_{s-}^1, s)\right] dL_t^b \\ &= \frac{1}{2}\int_0^t \left[F_x(X_{s-}+, s) - F_x(X_{s-}-, s)\right] 1_{\{X_{s-}=b(s)\}} dL_t^b,\end{aligned}$$

using Theorem 69 of [16]. On the other hand, the jump term in (A.25) can be written as

$$\begin{aligned}\text{(A.27)} \quad & \\ \sum_{0<s\le t} & \Bigg[F(X_s, s) - F(X_{s-}, s) - \Delta X_s F_x(X_{s-}-, s) 1_{\{X_{s-}\le b(s)\}} \\ & \quad - \Delta X_s F_x(s, X_{s-}+) 1_{\{X_{s-}>b(s)\}}\Bigg] \\ &= \sum_{0<s\le t} \left[F(X_s, s) - F(X_{s-}, s) - \frac{1}{2}\Delta X_s \left[F_x(X_{s-}-, s) + F_x(X_{s-}+, s)\right]\right].\end{aligned}$$

This completes the proof. □

Acknowledgments. I would like to thank the two referees and the anonymous AE for their insightful comments. I also would like to express my gratitude to Farid Aitsahlia, Tom Bielecki, Masahiko Egami, Sebastian Jaimungal, Sergei Levendorskii, Mihai Sirbu, Mete Soner, Hao Xing, Virginia R. Young, Thaleia Zariphopoulou, and Gordan Zitkovic for their feedback.

REFERENCES

[1] L. Alili and A. E. Kyprianou, *Some remarks on first passage of Lévy processes, the American put and pasting principles*, Ann. Appl. Probab., 15 (2005), pp. 2062–2080.

[2] L. H. R. Alvarez, *Solving optimal stopping problems of linear diffusions by applying convolution approximations*, Math. Methods Oper. Res., 53 (2001), pp. 89–99.

[3] E. Bayraktar, *On the Perpetual American Put Options for Level Dependent Volatility Models with Jumps*, Technical report, University of Michigan, Ann Arbor, MI, 2008; available online from http://arxiv.org/pdf/math/0703538.

[4] E. Bayraktar, S. Dayanik, and I. Karatzas, *Adaptive Poisson disorder problem*, Ann. Appl. Probab., 16 (2006), pp. 1190–1261.

[5] B. Bouchard, N. El Karoui, and N. Touzi, *Maturity randomization for stochastic control problems*, Ann. Appl. Probab., 15 (2005), pp. 2575–2605.

[6] P. Carr, *Randomization and the American put*, Rev. Financ. Stud., 11 (1998), pp. 597–626.

[7] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice–Hall, Englewood Cliffs, NJ, 1964.

[8] A. Friedman, *Stochastic Differential Equations and Applications*, Dover, Mineola, NY, 2006.

[9] H. Ishii, *On uniqueness and existence of viscosity solutions of fully nonlinear second-order elliptic PDEs*, Comm. Pure Appl. Math., 42 (1989), pp. 15–45.

[10] I. Karatzas and S. E. Shreve, *Methods of Mathematical Finance*, Appl. Math. (N.Y.) 39, Springer-Verlag, New York, 1998.

[11] E. Mordecki and P. Salminen, *Optimal stopping of Hunt and Lévy processes*, Stochastics, 79 (2007), pp. 233–251.

[12] G. Peskir, *A change-of-variable formula with local time on curves*, J. Theoret. Probab., 18 (2005), pp. 499–535.

[13] G. Peskir and A. Shiryaev, *Optimal Stopping and Free-Boundary Problems*, Lectures Math. ETH Zürich, Birkhäuser Verlag, Basel, 2006.

[14] H. Pham, *Optimal stopping, free boundary, and American option in a jump-diffusion model*, Appl. Math. Optim., 35 (1997), pp. 145–164.

[15] H. Pham, *Optimal stopping of controlled jump diffusion processes: A viscosity solution approach*, J. Math. Systems Estim. Control, 8 (1998).

[16] P. E. Protter, *Stochastic Integration and Differential Equations*, Stoch. Model. Appl. Probab. 21, Springer-Verlag, Berlin, 2005.

[17] R. T. Rockafellar, *Convex Analysis*, Princeton Landmarks Math., Princeton University Press, Princeton, NJ, 1997.

[18] P. Wilmott, S. Howison, and J. Dewynne, *The Mathematics of Financial Derivatives: A Student Introduction*, Cambridge University Press, Cambridge, UK, 1995.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 2, pp. 573–593

© 2009 Society for Industrial and Applied Mathematics

DYNAMIC OUTPUT FEEDBACK CONTROL OF DISCRETE-TIME MARKOV JUMP LINEAR SYSTEMS THROUGH LINEAR MATRIX INEQUALITIES[*]

JOSÉ C. GEROMEL[†], ALIM P. C. GONÇALVES[†], AND ANDRÉ R. FIORAVANTI[‡]

Abstract. This paper addresses the $\mathcal{H}_2$ and $\mathcal{H}_\infty$ dynamic output feedback control design problems of discrete-time Markov jump linear systems. Under the mode-dependent assumption, which means that the Markov parameters are available for feedback, the main contribution is the complete characterization of all full order proper Markov jump linear controllers such that the $\mathcal{H}_2$ or $\mathcal{H}_\infty$ norm of the closed loop system remains bounded by a given prespecified level, yielding the global solution to the corresponding mode-dependent optimal control design problem, expressed in terms of pure linear matrix inequalities. Some academic examples are solved for illustration and comparison. As a more consequent practical application, the networked control of a vehicle platoon using measurement signals transmitted in a Markov channel, as initially proposed in [P. Seiler and R. Sengupta, *IEEE Trans. Automat. Control*, 50 (2005), pp. 356–364], is considered.

Key words. linear systems, discrete-time systems, stochastic systems, Markov jump linear systems, linear matrix inequalities

AMS subject classifications. 93C05, 93C55, 93E03, 93E25

DOI. 10.1137/080715494

1. Introduction.

In recent years, parameter-dependent dynamic systems have received a great amount of attention due to their flexibility to represent with precision real world situations with practical appeal. In this framework, linear parameter-varying and gain scheduling design problems appeared in the deterministic and stochastic contexts. The latter class is composed of control systems where the open loop model presents sudden changes on their structures or parameters, which being modeled as Markovian processes become decisive for the increasing interest in the so-called Markov jump linear systems (MJLS) in both continuous- and discrete-time domains. An important assumption to consider for MJLS design is if the Markov chain state, often called mode, is available or not to the controller at every instant of time. Based on that information the design is said to be either mode-dependent or mode-independent, respectively. In this paper only the first case is considered for the following main reasons: First, in many practical situations, the system parameters are measurable; see [9], [22], [23], and [24]. Second, as a limitation of the proposed design method based on linear matrix inequalities (LMIs), only full order mode-dependent linear controllers may be handled without introducing any kind of conservatism. The mode-independent version of the output feedback control problem needs further research effort towards its complete solution.

One of the first works in the literature dealing with this class of models was presented in [1]. After that, a large number of theory and design procedures have been

[*] Received by the editors February 12, 2008; accepted for publication (in revised form) September 24, 2008; published electronically February 13, 2009. This research was supported by "Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)" and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Brazil.

http://www.siam.org/journals/sicon/48-2/71549.html

[†] DSCE/School of Electrical and Computer Engineering, UNICAMP, CP 6101, 13081 - 970 Campinas, SP, Brazil (geromel@dsce.fee.unicamp.br, alimped@dsce.fee.unicamp.br).

[‡] Institut National de Recherche en Informatique et en Automatique, Domaine de Voluceau, BP 105, 78153 Le Chesnay Cedex, France (andre.fioravanti@inria.fr).

developed in order to extend the concepts of the deterministic systems to this special class, namely stability concepts and testable conditions [6], [19], [17]; optimal state feedback control [18]; state feedback $\mathcal{H}_2$ optimization via convex programming [5]; state feedback $\mathcal{H}_2$ optimization via LMIs [13]; state feedback $\mathcal{H}_\infty$ optimization and robustness via LMIs [8], [14], [12]; state feedback $\mathcal{H}_\infty$ via Riccati equations [2]; and $\mathcal{H}_2$ filtering [7].

A problem of theoretical and practical importance in this area is the dynamic output feedback design. For the continuous-time case, many results are available in the literature related to the complete solution of the associated $\mathcal{H}_2$ and $\mathcal{H}_\infty$ problems; see [10] and, more recently, [20]. However, the same is not true for discrete-time systems for which only a few results are available; see [9], [22], and [24]. Indeed, in [9] an important result is reported, extending to MJLS the validity of the separation principle in the case of $\mathcal{H}_2$ norm and strictly proper linear controllers. The output feedback control problem with $\mathcal{H}_\infty$ criterion has been treated in [22] but restricting the attention to strictly proper linear controllers and, in addition, treating exclusively a very particular Markov chain characterized by the transition probability matrix with identical rows. Finally, [24] proposes handling the problem by a relaxation technique applied to bilinear matrix inequalities (BMIs).

In this paper, both $\mathcal{H}_2$ and $\mathcal{H}_\infty$ are considered, and, contrary to what has been done in [22], we do not make any assumption about the probability transitions of the Markov chain. We believe that the present paper innovates in the following directions:

- The set of all full order, proper, and mode-dependent Markov jump linear controllers imposing a prespecified $\mathcal{H}_2$ or $\mathcal{H}_\infty$ norm level to the controlled output of the closed loop system is provided. As a consequence, from the solution of convex problems expressed in terms of pure LMIs, the global optimal $\mathcal{H}_2$ or $\mathcal{H}_\infty$ controllers of this class are determined in only one shot, avoiding an iterative process and convergence difficulties to get the global solution [3].
- The controllers are parameterized by LMIs whose dimensions depend upon the dimension of the open loop system state variable and not on the number of modes of the Markov chain. This contributes decisively to decrease the computational burden involved.

The paper is organized as follows. In the next section, classical results such as stability and $\mathcal{H}_2$ and $\mathcal{H}_\infty$ norm calculations using LMIs are presented. In the same section a one-to-one change of variables used throughout for the linearization of the previously mentioned norms nonlinear dependence with respect to the matrices of the controller state space realization is also introduced. In section 3 the $\mathcal{H}_2$ norm control design problem is solved. In section 4 the same is done for the $\mathcal{H}_\infty$ norm. Notice that in the last two sections all results are necessary and sufficient for the class of controllers (full order, proper, linear, and mode-dependent) considered. Section 5 is devoted to presenting a practical application of the theoretical results obtained so far. It consists of networked control of a vehicle platoon using measurement signals transmitted in a Markov channel modelled in [22] but, in our opinion, with a more realistic, from the practical viewpoint, transition probability matrix. Section 6 presents the main conclusions of the paper and brief considerations on further works. Finally, Appendix I in section 7 is used to introduce some mathematical properties on matrix inequalities, and in Appendix II in section 8 the detailed proofs of the two main results are provided.

The notation used throughout is standard. Capital letters denote matrices and small letters denote vectors. For scalars, small Greek letters are used. For real matrices or vectors, ($'$) indicates transpose. For square matrices, $\mathrm{Tr}(X)$ denotes the trace function of X being equal to the sum of its eigenvalues and, for the sake of easing the notation of partitioned symmetric matrices, the symbol $(\bullet)$ denotes generically each of its symmetric blocks. The set of natural numbers is denoted by $\mathbb{N}$, while $\mathbb{K} = \{1, \ldots, N\}$. The unitary simplex in $\mathbb{R}^N$ composed of all nonnegative vectors $\mu \in \mathbb{R}^N$ such that $\mu_1 + \cdots + \mu_N = 1$ is denoted by Λ. Given N^2 nonnegative real numbers p_{ij} satisfying $p_{i1} + \cdots + p_{iN} = 1$ for all $i \in \mathbb{K}$ and N positive definite matrices $X_j \in \mathbb{R}^{n \times n}$ for all $j \in \mathbb{K}$, the convex combination of these matrices with weights p_{ij} is denoted by $X_{pi} = \sum_{j=1}^{N} p_{ij} X_j$ for all $i \in \mathbb{K}$. Similarly, for positive definite matrices, the inverse of the convex combination of inverses is denoted as

$$X_{qi} = \left(\sum_{j=1}^{N} p_{ij} X_j^{-1} \right)^{-1} . \tag{1}$$

Clearly, X_{pi} depends linearly on matrices $X_1, \ldots, X_N$, while the dependence of X_{qi} with respect to the same matrices is highly nonlinear. The same mathematical manipulations are adopted for positive definite matrices depending on two indices $i, j \in \mathbb{K} \times \mathbb{K}$. The symbol $\mathcal{E}\{\cdot\}$ denotes mathematical expectation of $\{\cdot\}$. For any stochastic signal $\xi(k)$, defined in the discrete-time domain $k \in \mathbb{N}$, the quantity $\|\xi\|_2^2 = \sum_{k=0}^{\infty} \mathcal{E}\{\xi(k)'\xi(k)\}$ is its squared norm. The class of all signals $\xi(k) \in \mathbb{R}^r$, $k \in \mathbb{N}$ such that $\|\xi\|_2^2$ is finite is denoted $\mathcal{L}_2^r$.

2. Problem formulation and basic results. A discrete-time MJLS is described by the following stochastic equations:

$$\mathbb{G} : \left\{ \begin{array}{rcl} x(k+1) & = & A(\theta_k)x(k) + B(\theta_k)u(k) + J(\theta_k)w(k), \\ z(k) & = & C_z(\theta_k)x(k) + D_z(\theta_k)u(k) + E_z(\theta_k)w(k), \\ y(k) & = & C_y(\theta_k)x(k) + E_y(\theta_k)w(k), \end{array} \right. \tag{2}$$

where $x(k) \in \mathbb{R}^n$ is the state, $u(k) \in \mathbb{R}^m$ is the control, $w(k) \in \mathbb{R}^p$ is the external perturbation, $z(k) \in \mathbb{R}^r$ is the controlled output, and $y(k) \in \mathbb{R}^q$ is the measured output. The state space matrices (2) depend upon a Markov chain taking values in the finite set $\mathbb{K}$ with the associated transition probability matrix given by $p_{ij} = \text{Prob}(\theta_{k+1} = j \mid \theta_k = i)$, which clearly satisfies the normalized constraints $p_{ij} \geq 0$ and $\sum_{j=1}^{N} p_{ij} = 1$ for each $i \in \mathbb{K}$. To ease the presentation, the notations $A(\theta_k) := A_i$, $B(\theta_k) := B_i$, $J(\theta_k) := J_i$, $C_z(\theta_k) := C_{zi}$, $D_z(\theta_k) := D_{zi}$, $E_z(\theta_k) := E_{zi}$, $C_y(\theta_k) := C_{yi}$, and $E_y(\theta_k) := E_{yi}$ whenever $\theta_k = i \in \mathbb{K}$ are adopted. The first important concept related to the model (2) is stability. In the context of MJLS, there are several equivalent forms to define stability as summarized in the next definition; see [19].

Definition 2.1. *Consider the model* (2) *with null control* $u(k) \equiv 0$*, null external input* $w(k) \equiv 0$ *for all* $k \in \mathbb{N}$*, and initial conditions* $x(0) = x_0 \in \mathbb{R}^n$*,* $\theta_0 \in \mathbb{K}$*. The system* $\mathbb{G}$ *is*

(a) *mean square stable if for every initial state* (x_0, θ_0)

$$\lim_{k \to \infty} \mathcal{E}\{x(k)'x(k) | x_0, \theta_0\} = 0, \tag{3}$$

(b) *stochastically stable if for every initial state* (x_0, θ_0)

$$\mathcal{E}\left\{ \sum_{k=0}^{\infty} x(k)'x(k) | x_0, \theta_0 \right\} < \infty, \tag{4}$$

(c) *exponentially mean square stable if for every initial state* (x_0, θ_0) *there exist constants* $0 < \alpha < 1$ *and* $\beta > 0$ *such that for all* $k \in \mathbb{N}$

$$\mathcal{E}\{x(k)'x(k)|x_0, \theta_0\} < \beta\alpha^k x_0' x_0. \tag{5}$$

It has been shown in [19] that the above definitions of stability are actually equivalent for an MJLS being referred to as second-moment stability (SMS). The next proposition [6], [18] presents a method to check stability from the existence of a positive definite solution of a set of coupled Lyapunov-like inequalities.

LEMMA 2.2. *The following statements are equivalent:*

(i) *System* $\mathbb{G}$ *is stable.*

(ii) *There exist* $P_i = P_i' > 0$ *such that*

$$A_i' P_{pi} A_i - P_i < 0 \tag{6}$$

for all $i \in \mathbb{K}$.

Although the inequalities (6) are already in the form of LMIs, as will be clear in what follows, they present some difficulties to be circumvented. Basically, the main technical difficulty is the way the summations that take the jump probabilities into account appear in the inequalities, involving the system dynamics and preventing one from using the standard transformations available in the literature to linearize the controller formulas. For the sake of comparison, recall that in continuous-time MJLS these summations appear as additive terms to the standard inequalities and never involve products with the dynamic matrices [10].

The following lemma provides an alternative characterization of stability for discrete-time MJLS. As will be shown later, these inequalities are more appropriate for dynamic output feedback control design.

LEMMA 2.3. *The following statements are equivalent:*

(i) *System* $\mathbb{G}$ *is stable.*

(ii) *There exist* $P_i = P_i' > 0$ *such that*

$$\begin{bmatrix} P_i & A_i' \\ A_i & P_{pi}^{-1} \end{bmatrix} > 0 \tag{7}$$

for all $i \in \mathbb{K}$.

The nonlinear inequalities in Lemma 2.3 have several formal advantages over the linear ones appearing in Lemma 2.2. The notation used in Lemma 2.3 puts into evidence that the inequalities required for testing stability involve only matrices of the same index $i \in \mathbb{K}$ and the coupling between the indices can be dealt with by the linear equality constraint $P_{pi} = \sum_{j=1}^{N} p_{ij} P_j$, which does not involve the dynamic system matrices A_i, for all $i \in \mathbb{K}$. Such a feature will be of extreme importance in deriving simple and effective formulas for an adequate parameterization of the controller state space matrices. The next definition is the generalization of the $\mathcal{H}_2$ norm from linear time invariant (LTI) systems to the stochastic Markovian jump case under consideration.

DEFINITION 2.4. *The* $\mathcal{H}_2$ *norm of a stable system* $\mathbb{G}$ *from the input* w *to the output* z *is given by*

$$\|\mathbb{G}\|_2^2 := \sum_{i=1}^{N} \sum_{s=1}^{p} \mu_i \|z^{s,i}\|_2^2, \tag{8}$$

where $\mu_i = \mathrm{Prob}(\theta_0 = i)$ *and* $z^{s,i}$ *represents the controlled output* $z(0), z(1), \ldots$ *obtained from the input* $w(k) = e_s\delta(k)$*, where* $e_s \in \mathbb{R}^p$ *is the sth column of the* $p \times p$ *identity matrix,* $\delta(k)$ *is the discrete impulse function,* $x(0) = 0$*, and* $\theta_0 = i \in \mathbb{K}$*.*

It is interesting to observe that in the deterministic case characterized by $N = 1$ the previous definition reduces to the usual $\mathcal{H}_2$ norm of the LTI discrete-time system $\mathbb{G}$. Moreover, in the general case, we have

$$\begin{aligned} \|\mathbb{G}\|_2^2 &= \sum_{i=1}^{N}\sum_{s=1}^{p} \mu_i \|z^{s,i}\|_2^2 \\ &\le \max_{\mu\in\Lambda} \sum_{i=1}^{N}\sum_{s=1}^{p} \mu_i \|z^{s,i}\|_2^2 \\ &\le \sup_{\theta_0\in\mathbb{K}} \sum_{s=1}^{p} \|z^{s,\theta_0}\|_2^2, \end{aligned} \tag{9}$$

which shows that for an adequate choice of the initial probabilities μ_i, $i \in \mathbb{K}$, the $\mathcal{H}_2$ norm of $\mathbb{G}$ equals the worst case norm (9) corresponding to the fact that no information about the initial state $\theta_0 \in \mathbb{K}$ is available. Of course, to determine the worst case norm the initial probability should be included in the optimization process as an additional variable to be determined. This point will be addressed in what follows with more details. For the moment, the next proposition shows how the $\mathcal{H}_2$ norm can be calculated [5].

LEMMA 2.5. *Assume that* $\mathbb{G}$ *is stable. The* $\mathcal{H}_2$ *norm of system* $\mathbb{G}$ *defined in* (8) *is given by*

$$\|\mathbb{G}\|_2^2 = \sum_{i=1}^{N} \mu_i \,\mathrm{Tr}\left(J_i' P_{pi} J_i + E_{zi}' E_{zi}\right), \tag{10}$$

where $P_i = P_i' > 0$ *solve* $P_i = A_i' P_{pi} A_i + C_{zi}' C_{zi}$ *for all* $i \in \mathbb{K}$*.*

From this result, there is no difficulty in calculating $\|\mathbb{G}\|_2^2$ using a standard LMI solver. The key observation is that if the matrix equality is replaced by inequalities, then the right-hand side of (10) is still an upper bound whose minimization provides $\|\mathbb{G}\|_2^2$. Hence, we have

$$\|\mathbb{G}\|_2^2 = \inf_{(W_i,P_i)\in\Phi} \sum_{i=1}^{N} \mu_i \,\mathrm{Tr}\left(W_i\right), \tag{11}$$

where Φ is the set of all matrices (W_i, P_i) for $i \in \mathbb{K}$ such that the LMIs

$$\begin{bmatrix} W_i & J_i' P_{pi} & E_{zi}' \\ \bullet & P_{pi} & 0 \\ \bullet & \bullet & I \end{bmatrix} > 0 \tag{12}$$

and

$$\begin{bmatrix} P_i & A_i' P_{pi} & C_{zi}' \\ \bullet & P_{pi} & 0 \\ \bullet & \bullet & I \end{bmatrix} > 0 \tag{13}$$

are satisfied for all $i \in \mathbb{K}$. From the numerical point of view the determination of $\|\mathbb{G}\|_2^2$ using LMIs appears to be adequate and efficient. Indeed, we have to handle $2N$ LMIs with $2N$ matrix variables and the coupling terms $P_{pi} = \sum_{j=1}^{N} p_{ij} P_j$ for all $i \in \mathbb{K}$. In addition, the global optimal solution of the convex programming problem (11) is calculated in only one shot.

Problem (11) can be slightly modified for the determination of the worst case norm defined in (9). Actually, keeping in mind that the maximum of N numbers equals the minimum upper bound we have

$$\begin{aligned}\sup_{\theta_0 \in \mathbb{N}} \sum_{s=1}^{p} \|z^{s,\theta_0}\|_2^2 &= \max_{\mu \in \Lambda} \inf_{(W_i,P_i) \in \Phi} \sum_{i=1}^{N} \mu_i \operatorname{Tr}(W_i) \\ &= \inf_{(W_i,P_i) \in \Phi} \max_{\mu \in \Lambda} \sum_{i=1}^{N} \mu_i \operatorname{Tr}(W_i) \\ &= \inf_{\sigma,(W_i,P_i) \in \Phi} \{\sigma \ : \ \operatorname{Tr}(W_i) < \sigma \ \forall i \in \mathbb{K}\}, \end{aligned} \tag{14}$$

which shows that the worst case norm can be determined with a little additional effort corresponding to increasing the number of variables by only one and the number of LMIs by N. It is important to stress that the second equality in (14) is due to the fact that all constraints are convex and the objective function is convex (linear) with respect to the minimization variables and concave (linear) with respect to the maximization variables. Finally, the last equality follows from duality relations.

We now move our attention to the $\mathcal{H}_\infty$ norm of the MJLS $\mathbb{G}$ with state space realization given in (2). The formal definition of this important concept is as follows.

Definition 2.6. *The $\mathcal{H}_\infty$ norm of a stable system $\mathbb{G}$ from the input w to the output z is given by*

$$\|\mathbb{G}\|_\infty^2 = \sup_{0 \neq w \in \mathcal{L}_2^p,\, \theta_0 \in \mathbb{K}} \frac{\|z\|_2^2}{\|w\|_2^2}. \tag{15}$$

Once again, it is interesting to observe that in the deterministic case characterized by $N = 1$ the previous definition reduces to the usual $\mathcal{H}_\infty$ norm of the LTI discrete-time system $\mathbb{G}$. The next lemma shows how the $\mathcal{H}_\infty$ norm of the MJLS (2) can be calculated [4], [21].

Lemma 2.7. *The system $\mathbb{G}$ is stable and satisfies the norm constraint $\|\mathbb{G}\|_\infty^2 < \gamma$ if and only if there exist matrices $P_i = P_i' > 0$ such that*

$$\begin{bmatrix} A_i & J_i \\ C_{zi} & E_{zi} \end{bmatrix}' \begin{bmatrix} P_{pi} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} A_i & J_i \\ C_{zi} & E_{zi} \end{bmatrix} - \begin{bmatrix} P_i & 0 \\ 0 & \gamma I \end{bmatrix} < 0 \tag{16}$$

holds for all $i \in \mathbb{K}$.

Lemma 2.7 is a bounded real lemma for the MJLS (2). It can be obtained from the conditions derived in [4]; see also [22], [21] and the references therein. Notice that, as we have already mentioned, (16) reduces to the deterministic $\mathcal{H}_\infty$ norm condition for $N = 1$ and to be feasible it requires the existence of positive definite matrices P_i such that $A_i' P_{pi} A_i - P_i + C_{zi}' C_{zi} < 0$ for all $i \in \mathbb{K}$. This is possible if and only if $\mathbb{G}$ is stable; see [19].

From this result, there is no difficulty in calculating the norm $\|\mathbb{G}\|_\infty^2$ from the optimal solution of a convex programming problem expressed by LMIs. Indeed, applying the Schur complement to (16) it is seen that

$$\|\mathbb{G}\|_\infty^2 = \inf_{(\gamma, P_i)\in\Psi} \gamma, \tag{17}$$

where Ψ is the set of all positive definite matrices P_i and $\gamma \in \mathbb{R}$ such that the LMI

$$\begin{bmatrix} P_i & 0 & A_i' P_{pi} & C_{zi}' \\ \bullet & \gamma I & J_i' P_{pi} & E_{zi}' \\ \bullet & \bullet & P_{pi} & 0 \\ \bullet & \bullet & \bullet & I \end{bmatrix} > 0 \tag{18}$$

is satisfied for all $i \in \mathbb{K}$. As before, from the numerical point of view the same conclusions can be drawn. The determination of $\|\mathbb{G}\|_\infty^2$ using LMIs appears to be adequate and efficient. Indeed, we have to handle N LMIs with N matrix variables and the coupling terms $P_{pi} = \sum_{j=1}^N p_{ij} P_j$ for all $i \in \mathbb{K}$. In addition, the calculation of the global optimal solution of the convex programming problem (17) does not need iterations, and, consequently, no convergence condition has to be verified.

We are now in position to state the dynamic output feedback control design problems to be dealt with in the rest of this paper. Associated with (2) consider the full order proper mode-dependent Markov jump linear controller

$$\mathbb{C} : \begin{cases} x_c(k+1) &= A_c(\theta_k) x_c(k) + B_c(\theta_k) y(k), \\ u(k) &= C_c(\theta_k) x_c(k) + D_c(\theta_k) y(k), \end{cases} \tag{19}$$

where $x_c(k) \in \mathbb{R}^n$, $x_c(0) = 0$, and the matrices A_{ci}, B_{ci}, C_{ci}, and D_{ci} for all $i \in \mathbb{K}$ are of compatible dimensions. The goal is to determine these matrices in such a way that the $\mathcal{H}_2$ or the $\mathcal{H}_\infty$ norm of the closed loop system is minimized. Connecting the controller (19) to the system (2) the controlled output is given by

$$\mathbb{F} : \begin{cases} \tilde{x}(k+1) &= \tilde{A}(\theta_k)\tilde{x}(k) + \tilde{J}(\theta_k) w(k), \\ z(k) &= \tilde{C}(\theta_k)\tilde{x}(k) + \tilde{E}(\theta_k) w(k), \end{cases} \tag{20}$$

where the indicated matrices are

$$\tilde{A}_i := \begin{bmatrix} A_i + B_i D_{ci} C_{yi} & B_i C_{ci} \\ B_{ci} C_{yi} & A_{ci} \end{bmatrix}, \quad \tilde{J}_i := \begin{bmatrix} J_i + B_i D_{ci} E_{yi} \\ B_{ci} E_{yi} \end{bmatrix}, \tag{21}$$

$$\tilde{C}_i := \begin{bmatrix} C_{zi} + D_{zi} D_{ci} C_{yi} & D_{zi} C_{ci} \end{bmatrix}, \quad \tilde{E}_i := E_{zi} + D_{zi} D_{ci} E_{yi}; \tag{22}$$

hence, the problem to be solved is written in the final form

$$\min_{A_{ci}, B_{ci}, C_{ci}, D_{ci}} \|\mathbb{F}\|_h^2, \tag{23}$$

where $h = 2$ or $h = \infty$. It is important to make clear that the above formulation of the dynamic output feedback control design problem is highly nonconvex and difficult to solve; that is, in this form, it is not possible to calculate its global optimal solution. The reason is that the calculation of the objective function $\|\mathbb{F}\|_h^2$ depends upon a set of auxiliary variables (see problems (11) and (17)) which multiply the controller variables producing, consequently, a nonconvex problem. The way to circumvent this difficulty is to introduce a one-to-one change of variables able to linearize the nonlinear constraints to be handled.

From the previous determination of $\mathcal{H}_2$ and $\mathcal{H}_\infty$ norms, it can be seen that the full order controller $\mathbb{C}$ imposes a closed loop system $\mathbb{F}$ twice the order of the plant $\mathbb{G}$. Hence the $\mathcal{H}_2$ and $\mathcal{H}_\infty$ norm calculations need auxiliary symmetric matrices $\tilde{P}_i \in \mathbb{R}^{2n\times 2n}$ for all $i \in \mathbb{K}$. Accordingly, let $\tilde{P}_i$ be $2n \times 2n$ real matrices partitioned as follows:

$$\tag{24} \tilde{P}_i = \begin{bmatrix} X_i & U_i \\ U_i' & \hat{X}_i \end{bmatrix}, \quad \tilde{P}_i^{-1} = \begin{bmatrix} Y_i & V_i \\ V_i' & \hat{Y}_i \end{bmatrix}, \quad \tilde{T}_i = \begin{bmatrix} Y_i & I \\ V_i' & 0 \end{bmatrix},$$

where all blocks are $n \times n$ real matrices. It is immediately verified that

$$\tag{25} \tilde{T}_i'\tilde{P}_i\tilde{T}_i = \begin{bmatrix} Y_i & I \\ I & X_i \end{bmatrix}$$

for all $i \in \mathbb{K}$. It is a well known fact (see Appendix I in section 7 for a more detailed discussion) that if the matrix in (25) is constrained to be definite positive, then it is always possible to determine the matrix blocks in (24) in order to get $\tilde{P}_i > 0$. Moreover, this can be accomplished even if matrix U_i or V_i for each $i \in \mathbb{K}$ is arbitrarily (nonsingular) fixed. Now, we proceed by considering $\tilde{P}_i > 0$ and adopting a similar reasoning to the convex combination of these matrices. From (24), the same partition yields

$$\tag{26} \tilde{P}_{pi} = \sum_{j=1}^{N} p_{ij}\tilde{P}_j = \begin{bmatrix} X_{pi} & U_{pi} \\ U_{pi}' & \hat{X}_{pi} \end{bmatrix},$$

and denoting

$$\tag{27} \tilde{P}_{pi}^{-1} = \begin{bmatrix} R_{1i} & R_{2i} \\ R_{2i}' & R_{3i} \end{bmatrix}, \quad \tilde{Q}_i = \begin{bmatrix} I & X_{pi} \\ 0 & U_{pi}' \end{bmatrix},$$

it is verified that

$$\tag{28} \tilde{Q}_i'\tilde{P}_{pi}^{-1}\tilde{Q}_i = \begin{bmatrix} R_{1i} & I \\ I & X_{pi} \end{bmatrix}.$$

It is important to stress that the four block matrices which define the inverse $\tilde{P}_{pi}^{-1}$ depend nonlinearly on the four block matrices of $\tilde{P}_{pi}$. However, since $R_{1i}^{-1} = X_{pi} - U_{pi}\hat{X}_{pi}^{-1}U_{pi}'$, setting U_i such that $U_i = -\hat{X}_i$ the partitioned matrix in (28) becomes

$$\tag{29} \tilde{Q}_i'\tilde{P}_{pi}^{-1}\tilde{Q}_i = \begin{bmatrix} (X_{pi} + U_{pi})^{-1} & I \\ I & X_{pi} \end{bmatrix}.$$

From the above discussion, we mention again that the particular choice $U_i = -\hat{X}_i$ can be made without loss of generality and constrains matrix U_i to be symmetric and negative definite. Furthermore, (24) provides $U_i = -\hat{X}_i = Y_i^{-1} - X_i$, which enables us to rewrite (29) in the final form

$$\tag{30} \tilde{Q}_i'\tilde{P}_{pi}^{-1}\tilde{Q}_i = \begin{bmatrix} Y_{qi} & I \\ I & X_{pi} \end{bmatrix}.$$

Moreover, in the general case, that is, without the particular choice $U_i = -\hat{X}_i$, the equality $R_{1i} = Y_{qi}$ does not hold any longer, but matrices R_{1i} given by

$$\tag{31} R_{1i}^{-1} = X_{pi} - U_{pi}\hat{X}_{pi}^{-1}U_{pi}'$$

satisfy the inequalities

$$\begin{aligned} R_{1i}^{-1} &\geq \sum_{j=1}^{N} p_{ij}(X_j - U_j\hat{X}_j^{-1}U_j') \\ &\geq \sum_{j=1}^{N} p_{ij}Y_j^{-1} \\ &\geq Y_{qi}^{-1} \end{aligned} \tag{32}$$

for all $i \in \mathbb{K}$. The relations (30) and (32) are the key results to be used afterwards for dynamic output feedback control synthesis.

Additionally, the results to be presented in what follows are based on the linearization of the matrix inequalities involved in the norm calculations. Hence, let us introduce the following one-to-one change of variables:

$$\begin{bmatrix} A_{ci} & B_{ci} \\ C_{ci} & D_{ci} \end{bmatrix} = \begin{bmatrix} U_{pi} & X_{pi}B_i \\ 0 & I \end{bmatrix}^{-1} \begin{bmatrix} M_i - X_{pi}A_iY_i & F_i \\ L_i & K_i \end{bmatrix} \begin{bmatrix} V_i' & 0 \\ C_{yi}Y_i & I \end{bmatrix}^{-1}, \tag{33}$$

which from matrices (M_i, F_i, L_i, K_i) uniquely determine the dynamic output feedback controller matrices $(A_{ci}, B_{ci}, C_{ci}, D_{ci})$ and vice versa for each $i \in \mathbb{K}$. Indeed, notice that in (33) the inverses exist whenever matrices U_i and V_i are nonsingular for all $i \in \mathbb{K}$. The importance of this change of variables is that it allows one to convert the $\mathcal{H}_2$ and $\mathcal{H}_\infty$ output feedback control design problems stated before into convex programming problems expressed in terms of LMIs.

3. $\mathcal{H}_2$ mode-dependent control design. Based on the previous results our main purpose in this section is to calculate the global optimal solution of the $\mathcal{H}_2$ mode-dependent dynamic output feedback control design problem (11) which can be stated as

$$\inf \sum_{i=1}^{N} \mu_i \operatorname{Tr}(W_i), \tag{34}$$

where the infimum is taken with respect to the matrix variables $\tilde{P}_i$, W_i, A_{ci}, B_{ci}, C_{ci}, and D_{ci} for all $i \in \mathbb{K}$ satisfying the inequalities

$$\begin{bmatrix} W_i & \tilde{J}_i' & \tilde{E}_i' \\ \bullet & \tilde{P}_{pi}^{-1} & 0 \\ \bullet & \bullet & I \end{bmatrix} > 0 \tag{35}$$

and

$$\begin{bmatrix} \tilde{P}_i & \tilde{A}_i' & \tilde{C}_i' \\ \bullet & \tilde{P}_{pi}^{-1} & 0 \\ \bullet & \bullet & I \end{bmatrix} > 0, \tag{36}$$

where the closed loop system state space matrices are given in (21), (22) and $\tilde{P}_i$ is partitioned as indicated in (24) for each $i \in \mathbb{K}$. It is important to stress that these nonlinear matrix inequalities are expressed equivalently in terms of the inverse $\tilde{P}_{pi}^{-1}$, which is essential to get a linearized version from the change of variables introduced before.

LEMMA 3.1. *There exist a mode-dependent output feedback linear controller of the form* (19) *and symmetric matrices $W_i, \tilde{P}_i > 0$ satisfying the inequalities* (35) *for all $i \in \mathbb{K}$ if and only if there exist symmetric matrices W_i, X_i, Y_i, Z_{ij} and matrices F_i, K_i, H_i of compatible dimensions satisfying the LMIs*

$$\begin{bmatrix} W_i & J_i' + E_{yi}'K_i'B_i' & J_i'X_{pi} + E_{yi}'F_i' & E_{zi}' + E_{yi}'K_i'D_{zi}' \\ \bullet & H_i + H_i' - Z_{pi} & I & 0 \\ \bullet & \bullet & X_{pi} & 0 \\ \bullet & \bullet & \bullet & I \end{bmatrix} > 0 \tag{37}$$

and

$$\begin{bmatrix} Z_{ij} & H_i' \\ \bullet & Y_j \end{bmatrix} > 0 \tag{38}$$

for all $i, j \in \mathbb{K} \times \mathbb{K}$. Furthermore, whenever (37)–(38) *are satisfied, a suitable solution for* (35) *is provided by* (33) *with $U_i = Y_i^{-1} - X_i$ and $V_i = Y_i$ for all $i \in \mathbb{K}$.*

Proof. See Appendix II in section 8. □

As we can see in Appendix II in section 8, the proof of Lemma 3.1 strongly depends upon the inequality $Y_{qi} \geq R_{1i}$ and on the existence of a particular choice of matrices U_i for all $i \in \mathbb{K}$ such that the equality holds. In addition, it is to be noticed that the linear matrix function $H_i + H_i' - Z_{pi}$ appearing in the second element of the main diagonal of the LMI (37) is used to successfully linearize the nonlinear matrix function Y_{qi} (as already commented, recall the notation $Z_{pi} := \sum_{j=1}^{N} p_{ij} Z_{ij}$). This aspect is also present in the next lemma, where the remaining constraint needed to calculate the $\mathcal{H}_2$ norm of the closed loop system is treated.

LEMMA 3.2. *There exist a mode-dependent output feedback linear controller of the form* (19) *and symmetric matrices $\tilde{P}_i > 0$ satisfying the inequalities* (36) *for all $i \in \mathbb{K}$ if and only if there exist symmetric matrices X_i, Y_i, Z_{ij} and matrices M_i, L_i, F_i, K_i, H_i of compatible dimensions satisfying the LMIs*

$$\begin{bmatrix} Y_i & I & Y_iA_i' + L_i'B_i' & M_i' & Y_iC_{zi}' + L_i'D_{zi}' \\ \bullet & X_i & A_i' + C_{yi}'K_i'B_i' & A_i'X_{pi} + C_{yi}'F_i' & C_{zi}' + C_{yi}'K_i'D_{zi}' \\ \bullet & \bullet & H_i + H_i' - Z_{pi} & I & 0 \\ \bullet & \bullet & \bullet & X_{pi} & 0 \\ \bullet & \bullet & \bullet & \bullet & I \end{bmatrix} > 0 \tag{39}$$

and

$$\begin{bmatrix} Z_{ij} & H_i' \\ \bullet & Y_j \end{bmatrix} > 0 \tag{40}$$

for all $i, j \in \mathbb{K} \times \mathbb{K}$. Furthermore, whenever (39)–(40) *are satisfied, a suitable solution for* (36) *is provided by* (33) *with $U_i = Y_i^{-1} - X_i$ and $V_i = Y_i$ for all $i \in \mathbb{K}$.*

Proof. See Appendix II in section 8. □

We want to stress that this lemma has a very interesting aspect as far as the nonlinear nature of inequality (36) is concerned. From the adequate definition of matrices $\tilde{T}_i$ and $\tilde{Q}_i$ the mentioned inequality, which depends on two different variables

$\tilde{P}_i$ and $\tilde{P}_{pi}^{-1}$, has been successfully linearized. Furthermore, since both lemmas make use of the same change of variables the following result is immediate.

THEOREM 3.3. *There exists a mode-dependent output feedback linear controller of the form* (19) *such that* $\|\mathbb{F}\|_2^2 < \gamma$ *if and only if there exists a feasible solution of LMIs* (37), (38), *and* (39) *satisfying*

$$\sum_{i=1}^{N} \mu_i \operatorname{Tr}(W_i) < \gamma. \tag{41}$$

In the affirmative case, a suitable mode-dependent Markov jump linear output feedback controller is defined by the state space matrices A_{ci}, B_{ci}, C_{ci}, *and* D_{ci}, *provided in* (33) *with* $U_i = Y_i^{-1} - X_i$ *and* $V_i = Y_i$ *for all* $i \in \mathbb{K}$.

Proof. This theorem follows from the results of Lemmas 3.1 and 3.2 together with the fact that $\|\mathbb{F}\|_2^2 < \sum_{i=1}^{N} \mu_i \operatorname{Tr}(W_i)$ for all feasible solutions of LMIs (37), (38), and (39). □

The most important consequence of Theorem 3.3 is that the optimal global solution of the $\mathcal{H}_2$ control design problem (23) can be alternatively determined from

$$\inf_{\mathcal{X}\in\Omega} \sum_{i=1}^{N} \mu_i \operatorname{Tr}(W_i), \tag{42}$$

where $\mathcal{X} = (W_i, X_i, Y_i, Z_{ij}, M_i, F_i, L_i, K_i, H_i)$ for all $i, j \in \mathbb{K} \times \mathbb{K}$ are the matrix variables and Ω is the set of all feasible solutions of LMIs (37), (38), and (39). In other words, the mode-dependent output feedback design problem under consideration has been converted into a convex programming problem expressed in terms of LMIs, which enables the use of efficient numerical methods for its solution. Moreover, Theorem 3.3 admits another interpretation. It provides the set of all mode-dependent proper full order Markov jump linear controllers such that the closed loop system satisfies the constraint $\|\mathbb{F}\|_2^2 < \gamma$ for some prespecified $\mathcal{H}_2$ norm level $\gamma > 0$. To the best of our knowledge, a similar result was not available in the literature to date.

Finally, it is interesting to see that the worst case norm (14) can be treated with no additional difficulty. Actually, from duality, the convexity of the primal design problem (42) enables us to conclude that

$$\max_{\mu\in\Lambda} \inf_{\mathcal{X}\in\Omega} \sum_{i=1}^{N} \mu_i \operatorname{Tr}(W_i) = \inf_{\sigma,\mathcal{X}\in\Omega} \{\sigma \ : \ \operatorname{Tr}(W_i) < \sigma\}, \tag{43}$$

making clear that the right-hand side of (43) provides the optimal mode-dependent Markov jump linear output feedback controller associated with the worst case norm.

3.1. Example. The following example was borrowed from [15]. It consists of two masses coupled with a spring and a damper. The first mass is attached to a fixed end point through another spring. The problem is to control the position and velocity of the second mass by applying a horizontal force on it. It is assumed that the position and velocity of the first mass are measured and that this information is delivered to the controller through a Markovian channel, which can insert error into the transmitted information package. Furthermore, it is also assumed that the controller can detect but not correct each of these defected packages by means of an adequate protocol, in which case it discards them. The force is applied directly on the

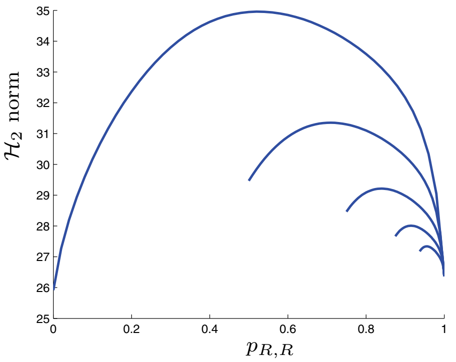


FIG. 1. *$\mathcal{H}_2$ norm versus $p_{R,R}$ for different values of κ.*

mass; i.e., it is not directly affected by the errors that may occur in the transmission channel.

The probability that a good package is received right after a good one is given by $p_{R,R}$, whereas the probability that a bad package is received after a bad one is given by $p_{L,L}$. Hence, the ratio between the mean length of good and bad package sequences is simply given by

$$\kappa = \frac{1 - p_{L,L}}{1 - p_{R,R}}. \tag{44}$$

Moreover, since $p_{L,L}$ and $p_{R,R}$ are probabilities for each fixed value of κ we must have $1 - \kappa^{-1} \leq p_{R,R} \leq 1$ and $p_{L,L} = (1 - \kappa) + \kappa p_{R,R}$. In practice, both probabilities $p_{R,R}$ and $p_{L,L}$ are expected to be high and $p_{R,R} > p_{L,L}$, which naturally yields $\kappa > 1$.

On the other hand, the state space realization of this continuous-time stochastic system is described as usual, with two modes ($N = 2$) and specific data given by

$$\left[\begin{array}{c|c|c} A_i & B_i & J_i \\ \hline C_{zi} & D_{zi} & E_{zi} \\ \hline C_{y1} & D_{y1} & E_{y1} \\ \hline C_{y2} & D_{y2} & E_{y2} \end{array}\right] = \left[\begin{array}{rrrr|r|r|r} 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ -30.0 & 10.0 & -0.36 & 0.36 & 0 & 0 & 0 \\ 5.0 & -5.0 & 0.18 & -0.18 & 1 & 0 & 0 \\ \hline 0 & 50.0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1.8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 & 0 & 0 & 0.1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0.125 \end{array}\right] \tag{45}$$

for $i \in \mathbb{K} = \{1, 2\}$. Notice that only the output matrices C_{yi} and E_{yi} for $i \in \mathbb{K}$ depend upon the Markov state in order to cope with the transmission error introduced by the Markov channel. The final discrete-time system of the form (2) is determined from discretization with sample time $T_s = 0.5$s and a zero order hold placed on each input. Considering $\mu_1 = 0$ and $\mu_2 = 1$, that is, the initial mode is the one corresponding to transmission without failure, the $\mathcal{H}_2$ mode-dependent proper controllers for a grid of the transition probability $p_{R,R}$ have been calculated, and in Figure 1 the closed loop system $\mathcal{H}_2$ performance provided by the optimal solution of problem (42) for $\kappa \in \{1, 2, 4, 8, 16\}$ is shown. Notice that for all values of $p_{R,R}$ as κ increases the minimum $\mathcal{H}_2$ norm decreases. Moreover, for the same κ, increasing $p_{R,R}$ the minimum $\mathcal{H}_2$ norm may also increase due to the fact that to keep κ constant the probability $p_{L,L}$ has to increase as well.

4. $\mathcal{H}_\infty$ mode-dependent control design. The main purpose of this section is to present similar results for the $\mathcal{H}_\infty$ norm. Connecting the full order linear controller $\mathbb{C}$ defined in (19) to the open loop system $\mathbb{G}$, the problem to be dealt with can be expressed as

$$\inf \gamma, \tag{46}$$

where the infimum is taken with respect to the scalar γ and the matrix variables $\tilde{P}_i$, A_{ci}, B_{ci}, C_{ci}, and D_{ci} for all $i \in \mathbb{K}$ satisfying the inequality

$$\begin{bmatrix} \tilde{P}_i & 0 & \tilde{A}_i' & \tilde{C}_i' \\ \bullet & \gamma I & \tilde{J}_i' & \tilde{E}_i' \\ \bullet & \bullet & \tilde{P}_{pi}^{-1} & 0 \\ \bullet & \bullet & \bullet & I \end{bmatrix} > 0, \tag{47}$$

where the closed loop system state space matrices are given in (21), (22) and $\tilde{P}_i$ is partitioned as indicated in (24) for each $i \in \mathbb{K}$. Comparing to the $\mathcal{H}_2$ mode-dependent control design analyzed in the previous section, the reader can verify that the same linearization strategy also works in the present context as summarized in the next theorem.

THEOREM 4.1. *There exist a mode-dependent output feedback linear controller of the form* (19)*, a scalar $\gamma > 0$, and symmetric matrices $\tilde{P}_i > 0$ satisfying the inequalities* (47) *for all $i \in \mathbb{K}$ if and only if there exist a scalar $\gamma > 0$, symmetric matrices X_i, Y_i, Z_{ij}, and matrices M_i, L_i, F_i, K_i, H_i of compatible dimensions satisfying the LMIs*

$$\begin{bmatrix} Y_i & I & 0 & Y_iA_i' + L_i'B_i' & M_i' & Y_iC_{zi}' + L_i'D_{zi}' \\ \bullet & X_i & 0 & A_i' + C_{yi}'K_i'B_i' & A_i'X_{pi} + C_{yi}'F_i' & C_{zi}' + C_{yi}'K_i'D_{zi}' \\ \bullet & \bullet & \gamma I & J_i' + E_{yi}'K_i'B_i' & J_i'X_{pi} + E_{yi}'F_i' & E_{zi}' + E_{yi}'K_i'D_{zi}' \\ \bullet & \bullet & \bullet & H_i + H_i' - Z_{pi} & I & 0 \\ \bullet & \bullet & \bullet & \bullet & X_{pi} & 0 \\ \bullet & \bullet & \bullet & \bullet & \bullet & I \end{bmatrix} > 0 \tag{48}$$

and

$$\begin{bmatrix} Z_{ij} & H_i' \\ \bullet & Y_j \end{bmatrix} > 0 \tag{49}$$

for all $i, j \in \mathbb{K} \times \mathbb{K}$. Furthermore, whenever (48)–(49) *are satisfied, a suitable solution for* (47) *is provided by* (33) *with $U_i = Y_i^{-1} - X_i$ and $V_i = Y_i$ for all $i \in \mathbb{K}$.*

Proof. The proof follows from similar arguments and mathematical relations already adopted in the proofs of Lemmas 3.1 and 3.2 as well; therefore it is being omitted. □

Defining the set of matrix variables $\mathcal{Y} = (X_i, Y_i, Z_{ij}, M_i, F_i, L_i, K_i, H_i)$ for all $i, j \in \mathbb{K} \times \mathbb{K}$ and the convex set Ξ of all feasible solutions of the LMIs (48) and (49),

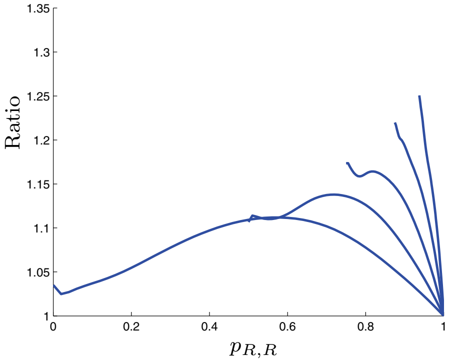


FIG. 2. *Performance ratio of optimal controllers for different values of* κ.

the optimal solution of problem (46) is determined from

$$\inf_{\gamma,\mathcal{Y}\in\Xi} \gamma. \tag{50}$$

Hence, as in the $\mathcal{H}_2$ norm case, the mode-dependent output feedback design problem under consideration has been converted into a convex programming problem expressed in terms of LMIs.

The result reported in Theorem 4.1 outperforms the previous results available in the literature dealing with this class of control design problems; see [22] and [24]. First, to the contrary of [22], where the very restrictive constraint on the transition probability matrix is imposed, namely, $p_{ij} = p_j$ for all $i, j \in \mathbb{K} \times \mathbb{K}$ and only strictly proper linear controllers are considered, we do not make any assumption about the structure of the transition probability matrix, and general proper controllers are designed. Moreover, comparing to [24], where an iterative method to solve BMIs is needed, here the problem is solved in one shot by any LMI solver.

4.1. Example. The data used in the example in section 3.1 is again considered to compare optimal performances of proper and strictly proper linear controllers. The $\mathcal{H}_\infty$ output feedback control problem (50) has also been solved in order to demonstrate how performance can be improved by allowing the controller to be proper instead of strictly proper. Figure 2 shows the performance ratio produced by strictly proper and proper optimal controllers for $\mathcal{H}_\infty$ norm optimization. As can be seen, the improvement is more expressive for bigger values of the transition probability $p_{R,R}$ and moderate values of $p_{L,L}$. This effect becomes more expressive whenever κ increases. Clearly, for $p_{R,R}$ and $p_{L,L}$ close to one the optimal controller becomes strictly proper.

5. Practical application. Consider the vehicle following problem described in [22]. Let x_0 denote the position of the leading car and x_i denote the position of the ith follower. The reference trajectory for the lead vehicle is denoted r_0 and the tracking error for the lead vehicle is $e_0 = r_0 - x_0$. The other vehicle spacing errors are $e_i = x_{i-1} - x_i - \delta_i$, where δ_i is the desired vehicle spacing. The control objective is to enforce all tracking errors e_i to zero.

Following [22], although the dynamic behavior for an individual vehicle is nonlinear, the use of a two-layered control scheme [16] allows us to consider a reasonable third order model for the vehicle dynamics according to

$$\frac{d}{dt}\begin{bmatrix} x_i(t) \\ v_i(t) \\ a_i(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & -\frac{1}{\tau} \end{bmatrix}\begin{bmatrix} x_i(t) \\ v_i(t) \\ a_i(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \frac{1}{\tau} \end{bmatrix} u_i(t), \tag{51}$$

where $x_i(t)$, $v_i(t)$, and $a_i(t)$ are the position, velocity, and acceleration of the ith vehicle, and $\tau = 100$ms is the time constant of the first order lag. The system is discretized with a $T_s = 20$ms sample rate, considering a zero order hold on the control input.

It is assumed that the measurement used to control the system is transmitted using a wireless network protocol such that errors can be detected but not corrected. At every sample time, a vehicle communicates its measurements to the network. If the received package is corrupted, then the controller discards it and waits for the next package. We assume that all communicated measurements from the vehicles are either received or corrupted and the Markov chain can be modelled with two modes $\{L, R\}$ for lost and received packages, respectively. It is further assumed that the statistics of the network are known so that we have both the probability of the next broadcasted package being correct after a good one is received $p_{R,R}$ and the probability of a package being lost by error after a bad one is received $p_{L,L}$. The data communicated through the network is modelled as follows:

$$\hat{y}_c(k) = \begin{cases} y_c(k) & \text{if } \theta(k) = R, \\ \emptyset & \text{if } \theta(k) = L, \end{cases} \tag{52}$$

where $\emptyset$ denotes a corrupted package of information and y_c is the vector of communicated measurements available for feedback. All cars in the platoon have on board sensors to capture the measurements from their particular motions. Those measurements are denoted as y_o, which leads to the following output available to the controller in every time instant:

$$y(k) := \begin{bmatrix} y_o(k) \\ \hat{y}_c(k) \end{bmatrix}. \tag{53}$$

As in [22], for the controller design a model with two cars is considered. However, the important difference from the approach proposed in [22] is that we are able to calculate with the results reported in this paper the optimal linear controller associated with any value of the transition probability matrix and not only for those satisfying $p_{R,R} = 1 - p_{L,L}$. Figure 3 shows the grid with the $\mathcal{H}_\infty$ norm of the controlled system for all possible values of the package loss rates.

It is interesting to notice, although it is not intuitive, that the $\mathcal{H}_\infty$ norm does not change very much with respect to $p_{R,R}$ whenever $p_{L,L}$ is kept constant. However, if one takes two points on Figure 3 with different $p_{R,R}$ and the same $p_{L,L}$, even though the $\mathcal{H}_\infty$ norm might be the same, the designed controllers are in general different. That cannot be accomplished by the control designed from [22] because it can be calculated only for systems such that $p_{L,L} = 1 - p_{R,R}$. Figure 3 shows spikes in some points due to lack of numerical precision.

Finally, we compare the Markov jump linear controller proposed by this design with the deterministic $\mathcal{H}_\infty$ optimal output feedback controller. For both systems a Monte Carlo simulation was run with 5×10^3 iterations. The probability matrix for the package error rates

$$\begin{bmatrix} p_{L,L} & p_{L,R} \\ p_{R,L} & p_{R,R} \end{bmatrix} = \begin{bmatrix} 0.92 & 0.08 \\ 0.02 & 0.98 \end{bmatrix} \tag{54}$$

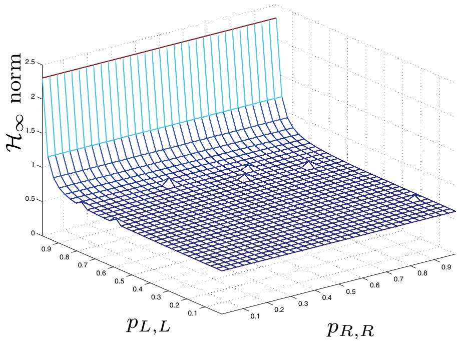


FIG. 3. *$\mathcal{H}_\infty$ norm versus package loss rate.*

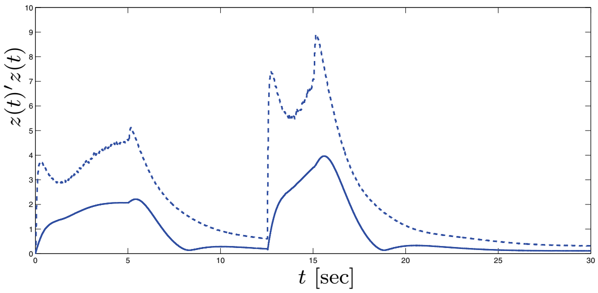


FIG. 4. *Mean square error for deterministic and Markovian controllers.*

has been considered, corresponding to the ratio $\kappa = 4$. The input has been calculated in order to impose that the lead vehicle should start with a constant acceleration of 3m/s^2 for 5 seconds, should remain with constant speed for the next 7.5 seconds, and should brake with an acceleration of -6m/s^2 for the remaining 2.5 seconds; after that, acceleration should be zero. The parameters used for the controlled output z, the spacing error e_1, and the weighted control effort were the same as those used in [22]. The mean square error can be seen in Figure 4 against $t = kT_s$ for $k \in [0,\, 1.5 \times 10^3]$, where the better performance of the Markovian controller (solid line) when compared to the deterministic one (dashed line) is clear.

6. Conclusion. In this paper a new approach to the dynamic output feedback control design of discrete-time MJLS is proposed. The set of all full order proper mode-dependent Markov jump linear output feedback controllers imposing on the closed loop system a prespecified $\mathcal{H}_2$ or $\mathcal{H}_\infty$ norm level is parameterized by convex constraints expressed by means of pure LMIs, likewise of what has been done in the continuous-time case [10]. The linear controllers are obtained without any additional constraints (as, for instance, to be strictly proper), a fact that is particularly important since Figure 2 makes clear the impact of this control structure as far as the $\mathcal{H}_\infty$ norm minimization is concerned. Furthermore, in that case, the performance is enhanced by the possibility of dealing with optimal control problems where the transition probability matrix is not restricted to having the same rows as considered in [22]. The controller is always obtained from the solution of a convex programming problem, assuming that it has access to the system modes for all $k \in \mathbb{N}$. This assumption is of practical appeal if one considers the application for networked control, where transmission protocols can easily include package error detection. The inclusion of the system statistics and the correspondent models to obtain the closed loop linear controller has a computational cost. Considering again the networked control with two possible modes as an example, the design of an optimal $\mathcal{H}_2$ controller will imply solving eight coupled LMIs against one for the deterministic case. Nonetheless, we believe it is a worthy cost to pay, given the better performance illustrated in Figure 4.

As a further research subject, we believe that the results reported in this paper can be applied to networked control problems where the Markov channel is described by more accurate models yielding higher order MJLS systems. Moreover, in our opinion, additional research efforts towards the generalization of the present results to the mode-independent versions of the $\mathcal{H}_2$ and $\mathcal{H}_\infty$ optimal control problems are of great theoretical and practical interest.

7. Appendix I. This appendix provides a series of basic results largely used in this paper. The following simple property involving square matrices has been applied with success to the solution of several problems in the literature to date (see [11] and the references therein).

LEMMA 7.1. *Consider $P = P' > 0$. The inequality*

$$G'P^{-1}G \geq G + G' - P \tag{55}$$

holds for any square matrix G of compatible dimensions.

Proof. It follows trivially from the fact that $(G - P)'P^{-1}(G - P) \geq 0$. □

Let a set of nonnegative real numbers p_{ij} for $i, j \in \mathbb{K} \times \mathbb{K}$ satisfying the normalization constraints

$$\sum_{j=1}^{N} p_{ij} = 1, \quad i \in \mathbb{K}, \tag{56}$$

and a set of positive definite matrices $X_j \in \mathbb{R}^{n \times n}$ for all $j \in \mathbb{K}$ be given. The convex combination of these matrices with weights p_{ij} is defined as

$$X_{pi} = \sum_{j=1}^{N} p_{ij} X_j, \quad i \in \mathbb{K}. \tag{57}$$

Similarly, the inverse of the convex combination of inverses of the same matrices is defined as

$$X_{qi} = \left(\sum_{j=1}^{N} p_{ij} X_j^{-1} \right)^{-1}, \quad i \in \mathbb{K}. \tag{58}$$

The next lemma gives a relationship between these two convex combinations that is exhaustively used in the present paper.

LEMMA 7.2. *Let $X_1, \ldots, X_N \in \mathbb{R}^{n \times n}$ be positive definite matrices. The inequality $X_{pi} \geq X_{qi}$ holds for all $i \in \mathbb{K}$.*

Proof. Notice that each inequality

$$\begin{bmatrix} X_j & I \\ I & X_j^{-1} \end{bmatrix} \geq 0, \quad j \in \mathbb{K}, \tag{59}$$

holds from the Schur complement and due to the fact that all matrices $X_1, \dots, X_N$ are positive definite. Hence, multiplying (59) by p_{ij} and summing up for all $j \in \mathbb{K}$, using (57) and (58) we obtain

$$\begin{bmatrix} X_{pi} & I \\ I & X_{qi}^{-1} \end{bmatrix} \geq 0, \quad i \in \mathbb{K}, \tag{60}$$

yielding, by the Schur complement, the desired result. □

Clearly, X_{pi} depends linearly upon matrices $X_1, \dots, X_N$, while the dependence of X_{qi} with respect to the same matrices is highly nonlinear. The importance of Lemma 7.2 is that it provides a linear upper bound to the nonlinear matrix function X_{qi} for all $i \in \mathbb{K}$.

We now move our attention to positive definite matrices with square partitions.

Lemma 7.3. *Let the symmetric matrices $X \in \mathbb{R}^{n\times n}$, $Y \in \mathbb{R}^{n\times n}$ such that*

$$\begin{bmatrix} Y & I \\ I & X \end{bmatrix} > 0 \tag{61}$$

be given. It is always possible to determine symmetric matrices $\hat{X} \in \mathbb{R}^{n\times n}$, $\hat{Y} \in \mathbb{R}^{n\times n}$ and matrices $U \in \mathbb{R}^{n\times n}$, $V \in \mathbb{R}^{n\times n}$ satisfying

$$\begin{bmatrix} X & U \\ U' & \hat{X} \end{bmatrix}^{-1} = \begin{bmatrix} Y & V \\ V' & \hat{Y} \end{bmatrix} > 0. \tag{62}$$

Proof. Take $U \in \mathbb{R}^{n\times n}$ nonsingular. The equality in (62) holds if and only if the following four matrix constraints simultaneously hold. First, $XY + UV' = I$, which gives $V = (I - YX)U'^{-1}$. Second, $XV + U\hat{Y} = 0$, yielding to the symmetric matrix $\hat{Y} = U^{-1}X(Y - X^{-1})XU'^{-1}$. Third, $U'Y + \hat{X}V' = 0$, which provides $\hat{X} = U'(X - Y^{-1})^{-1}U$. Finally, with these solutions we verify that

$$\begin{aligned} U'V + \hat{X}\hat{Y} &= U'(I - YX)U'^{-1} + U'(X - Y^{-1})^{-1}X(Y - X^{-1})XU'^{-1} \\ &= U'((I - YX) - (X - Y^{-1})^{-1}X(I - YX))U'^{-1} \\ &= U'(I - (X - Y^{-1})^{-1}X)(I - YX)U'^{-1} \\ &= U'(X - Y^{-1})^{-1}(-Y^{-1})(I - YX)U'^{-1} \\ &= I, \end{aligned} \tag{63}$$

which proves the equality in (62). On the other hand, since any feasible solution of (61) provides $I - YX$ nonsingular, the same is true for matrix V given above and, consequently, for the $2n \times 2n$ real matrix

$$T = \begin{bmatrix} Y & I \\ V' & 0 \end{bmatrix}. \tag{64}$$

Hence, from the fact that

$$T' \begin{bmatrix} X & U \\ U' & \hat{X} \end{bmatrix} T = \begin{bmatrix} Y & I \\ I & X \end{bmatrix} \tag{65}$$

the inequality in (62) follows immediately from (61). □

An interesting point arising from the proof of Lemma 7.3 is that (62) can be solved even though $U \in \mathbb{R}^{n\times n}$ is arbitrarily fixed as any nonsingular matrix. In other words, given symmetric matrices X and Y satisfying the positivity constraint (61), then, to solve (62), the nonsingular matrix U can be arbitrarily fixed without loss of generality. Simple verification puts into evidence that similar results hold for matrix $V \in \mathbb{R}^{n\times n}$.

8. Appendix II.

In this appendix we provide the proofs for Lemmas 3.1 and 3.2, which are central for the $\mathcal{H}_2$ output feedback control design problem.

8.1. Proof of Lemma 3.1.

For the necessity, assume that (35) holds. Partitioning $\tilde{P}_{pi}^{-1}$ as in (24) and multiplying (35) to the right by $\mathrm{diag}[I, \tilde{Q}_i, I]$ and to the left by its transpose we obtain

$$\begin{bmatrix} W_i & J_i' + E_{yi}'K_i'B_i' & J_i'X_{pi} + E_{yi}'F_i' & E_{zi}' + E_{yi}'K_i'D_{zi}' \\ \bullet & R_{1i} & I & 0 \\ \bullet & \bullet & X_{pi} & 0 \\ \bullet & \bullet & \bullet & I \end{bmatrix} > 0, \tag{66}$$

where $F_i = U_{pi}B_{ci} + X_{pi}B_iD_{ci}$ and $K_i = D_{ci}$. Taking into account that (32) implies $Y_{qi} \geq R_{1i}$, for $H_i = Y_{qi}$ and $Z_{ij} = Y_{qi}Y_j^{-1}Y_{qi} + \varepsilon I$ with $\varepsilon > 0$ we see that (38) is verified and we obtain

$$\begin{aligned} H_i + H_i' - Z_{pi} &= Y_{qi} - \varepsilon I \\ &\geq R_{1i} - \varepsilon I; \end{aligned} \tag{67}$$

hence, taking $\varepsilon > 0$ sufficiently small, inequality (66) implies that (37) holds and the claim follows.

For the sufficiency, assume that (37) and (38) hold. From (38) we have $Z_{ij} > H_i'Y_j^{-1}H_i$, and, consequently, multiplying these inequalities by p_{ij} and summing up for all $j \in \mathbb{K}$ we obtain

$$\begin{aligned} H_i + H_i' - Z_{pi} &= H_i + H_i' - \sum_{j=1}^{N} p_{ij}Z_{ij} \\ &\leq H_i + H_i' - H_i'Y_{qi}^{-1}H_i \\ &\leq Y_{qi} - (H_i - Y_{qi})'Y_{qi}^{-1}(H_i - Y_{qi}) \\ &\leq Y_{qi}, \end{aligned} \tag{68}$$

which implies that (37) remains valid if the diagonal term in the second column and row is replaced by Y_{qi} and consequently $X_{pi} > Y_{qi}^{-1} > 0$. Hence, imposing $U_i = Y_i^{-1} - X_i$ we get $V_i = Y_i$ and it is verified that the matrix $U_{pi} = Y_{qi}^{-1} - X_{pi}$ is nonsingular, which enable us to determine the matrices B_{ci} and D_{ci} from the change of variables (33). On the other hand, taking into account that this choice provides

$$\tilde{P}_{pi} = \begin{bmatrix} X_{pi} & Y_{qi}^{-1} - X_{pi} \\ \bullet & X_{pi} - Y_{qi}^{-1} \end{bmatrix} > 0 \tag{69}$$

it is immediately verified that (30) holds and that $R_{1i}^{-1} = Y_{qi}^{-1}$. The conclusion is that inequality (37) with the diagonal term in the second column and row replaced by Y_{qi}

can be rewritten as

$$\begin{bmatrix} W_i & \tilde{J}_i'\tilde{Q}_i & \tilde{E}_i' \\ \bullet & \tilde{Q}_i'\tilde{P}_{pi}^{-1}\tilde{Q}_i & 0 \\ \bullet & \bullet & I \end{bmatrix} > 0, \tag{70}$$

which multiplied to the right by $\mathrm{diag}[I, \tilde{Q}_i^{-1}, I]$ and to the left by its transpose provides the inequality (35), and the proof is concluded. □

8.2. Proof of Lemma 3.2. Along general lines, it follows the same pattern of the proof of Lemma 3.1.

For the necessity, assume that the inequality (36) holds. Partitioning $\tilde{P}_i$ and $\tilde{P}_{pi}^{-1}$ as indicated in (24) and (26), respectively, multiplying (36) to the right by $\mathrm{diag}[\tilde{T}_i, \tilde{Q}_i, I]$ and to left by its transpose, and adopting the inverse change of variables (33), we get the LMI (39) with R_{1i} at the place of $H_i + H_i' - Z_{pi}$ in the third row and third column block. As before, the necessity follows from (67), making $\varepsilon > 0$ sufficiently small.

The sufficiency follows from the particular choice of matrix $U_i = Y_i^{-1} - X_i$ implying from (24) that $V_i = Y_i$. Moreover, taking into account that this choice provides $\tilde{P}_{pi} > 0$ as in (30) and that $R_{1i} = Y_{qi} \geq H_i + H_i' - Z_{pi}$, the change of variables proposed enables us to get

$$\begin{bmatrix} \tilde{T}_i'\tilde{P}_i\tilde{T}_i & \tilde{T}_i'\tilde{A}_i'\tilde{Q}_i & \tilde{T}_i'\tilde{C}_i' \\ \bullet & \tilde{Q}_i'\tilde{P}_{pi}^{-1}\tilde{Q}_i & 0 \\ \bullet & \bullet & I \end{bmatrix} > 0, \tag{71}$$

which provides (36) after multiplication to the right by $\mathrm{diag}[\tilde{T}_i^{-1}, \tilde{Q}_i^{-1}, I]$ and to the left by its transpose. This concludes the proof of the proposed lemma. □

REFERENCES

[1] W. P. Blair, Jr., and D. D. Sworder, *Feedback control of a class of linear discrete systems with jump parameters and quadratic cost criteria*, Internat. J. Control, 21 (1975), pp. 833–841.

[2] E. K. Boukas and P. Shi, *H_∞ control for discrete-time linear systems with Markovian jumping parameters*, in Proceedings of the 36th IEEE Conference on Decision and Control, 1997, pp. 4134–4139.

[3] S. P. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, SIAM, Philadelphia, 1994.

[4] O. L. V. Costa and J. B. R. do Val, *Full information H_∞ control for discrete-time infinite Markov jump parameter systems*, J. Math. Anal. Appl., 202 (1996), pp. 578–603.

[5] O. L. V. Costa, J. B. R. do Val, and J. C. Geromel, *A convex programming approach to H_2 control of discrete-time Markovian linear systems*, Internat. J. Control, 66 (1997), pp. 557–579.

[6] O. L. V. Costa and M. D. Fragoso, *Stability results for discrete-time linear systems with Markovian jump parameters*, J. Math. Anal. Appl., 179 (1993), pp. 154–178.

[7] O. L. V. Costa and S. Guerra, *Stationary filter for linear minimum mean square error estimator of discrete-time Markovian jump systems*, IEEE Trans. Automat. Control, 47 (2002), pp. 1351–1356.

[8] O. L. V. Costa and R. P. Marques, *Mixed H_2/H_∞-control of discrete-time Markovian jump linear systems*, IEEE Trans. Automat. Control, 43 (1998), pp. 95–100.

[9] O. L. V. Costa and E. F. Tuesta, *H_2-control and the separation principle for discrete-time Markovian jump linear systems*, Math. Control Signals Systems, 16 (2004), pp. 320–350.

[10] D. P. de Farias, J. C. Geromel, J. B. R. do Val, and O. L. V. Costa, *Output feedback control of Markov jump linear systems in continuous-time*, IEEE Trans. Automat. Control, 45 (2000), pp. 944–949.

[11] M. C. de Oliveira, J. Bernussou, and J. C. Geromel, *A new discrete-time robust stability condition*, Systems Control Lett., 37 (1999), pp. 261–265.

[12] C. E. de Souza, *Mode-independent H_∞ control of discrete-time Markovian jump linear systems*, in Proceedings of the 16th IFAC World Congress, Prague, Czech Republic, 2005.

[13] J. B. R. do Val, J. C. Geromel, and A. P. C. Gonçalves, *The H_2-control for jump linear systems: Cluster observations of the Markov state*, Automatica J. IFAC, 38 (2002), pp. 343–349.

[14] L. El Ghaoui and M. A. Rami, *Robust state-feedback stabilization of jump linear systems via LMIs*, Internat. J. Robust Nonlinear Control, 6 (1996), pp. 1015–1022.

[15] A. R. Fioravanti, A. P. C. Gonçalves, and J. C. Geromel, *$\mathcal{H}_2$ and $\mathcal{H}_\infty$ filtering of discrete-time Markov jump linear systems through linear matrix inequalities*, in Proceedings of the 17th IFAC World Congress, Seoul, Korea, 2008, pp. 2681–2686.

[16] J. K. Hedrick, M. Tomikuza, and P. Varaiya, *Control issues in automated highway systems*, IEEE Control Syst. Mag., 14 (1994), pp. 21–32.

[17] Y. Ji and H. J. Chizeck, *Controllability, observability and discrete-time Markovian jump linear quadratic control*, Internat. J. Control, 48 (1988), pp. 481–498.

[18] Y. Ji and H. J. Chizeck, *Jump linear quadratic Gaussian control: Steady-state solution and testable conditions*, Control Theory Adv. Tech., 6 (1990), pp. 289–319.

[19] Y. Ji, H. J. Chizeck, X. Feng, and K. A. Loparo, *Stability and control of discrete-time jump linear systems*, Control Theory Adv. Tech., 7 (1991), pp. 247–270.

[20] L. Li and V. A. Ugrinovskii, *On necessary and sufficient conditions for $\mathcal{H}_\infty$ output feedback control of Markov jump linear systems*, IEEE Trans. Automat. Control, 52 (2007), pp. 1287–1292.

[21] P. Seiler and R. Sengupta, *A bounded real lemma for jump systems*, IEEE Trans. Automat. Control, 48 (2003), pp. 1651–1654.

[22] P. Seiler and R. Sengupta, *An H_∞ approach to networked control*, IEEE Trans. Automat. Control, 50 (2005), pp. 356–364.

[23] D. D. Sworder and R. O. Rogers, *An LQ-solution to a control problem associated with a solar thermal central receiver*, IEEE Trans. Automat. Control, 28 (1983), pp. 971–978.

[24] L. Xiao, A. Hassibi, and J. How, *Control with random communication delays via a discrete-time jump system approach*, in Proceedings of the American Control Conference, Chicago, IL, 2000, pp. 2199–2204.

SIAM J. Control Optim.
Vol. 48, No. 2, pp. 594–617

© 2009 Society for Industrial and Applied Mathematics

SMOOTH FIT PRINCIPLE FOR IMPULSE CONTROL OF MULTIDIMENSIONAL DIFFUSION PROCESSES*

XIN GUO† AND GUOLIANG WU‡

Abstract. Value functions of impulse control problems are known to satisfy quasi-variational inequalities (QVIs) [A. Bensoussan and J.-L. Lions, *Impulse Control and Quasivariational Inequalities*, Heyden & Son, Philadelphia, 1984; translation of *Contrôle Impulsionnel et Inéquations Quasi Variationnelles*, Gauthier-Villars, Paris, 1982]. This paper proves the smooth-fit C^1 property of the value function for multidimensional controlled diffusions, using a viscosity solution approach. We show by examples how to exploit this regularity property to derive explicitly optimal policy and value functions.

Key words. stochastic impulse control, viscosity solution, quasi-variational inequality, smooth fit, controlled diffusion

AMS subject classifications. 49J20, 49N25, 49N60

DOI. 10.1137/080716001

1. Introduction.

This paper considers the following impulse control problem for an n-dimensional diffusion process $X(t)$. In the absence of control, $X(t)$ is governed by an Itô's stochastic differential equation,

$$dX(t) = \mu(X(t))dt + \sigma(X(t))dW(t), \quad X(0) = x, \tag{1.1}$$

where W is a standard Brownian motion in a filtered probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If a control policy $V = (\tau_1, \xi_1; \tau_2, \xi_2; \ldots)$ is adopted, then $X(t)$ evolves as

$$dX(t) = \mu(X(t-))dt + \sigma(X(t-))dW(t) + \sum_i \delta(t - \tau_i)\xi_i, \tag{1.2}$$

where $\delta(\cdot)$ denotes the Dirac delta function. Here the control $V = (\tau_1, \xi_1; \tau_2, \xi_2; \ldots)$ is of an impulse type such that $\tau_1, \tau_2, \ldots$ is an increasing sequence of stopping times with respect to $\mathcal{F}_t$ (the natural filtration generated by W), and ξ_i is an $\mathbb{R}^n$-valued, $\mathcal{F}_{\tau_i}$-measurable random variable.

The problem is to choose an appropriate impulse control $(\tau_1, \xi_1; \tau_2, \xi_2; \ldots)$ so that the following objective function is minimized:

$$\mathbb{E}_x \left(\int_0^\infty e^{-rt} f(X(t))dt + \sum_{i=1}^\infty e^{-r\tau_i} B(\xi_i) \right). \tag{1.3}$$

Here f is a running cost function, B is a transaction cost function, and $r > 0$ is a discount factor.

This multidimensional control problem has been proposed and studied in various forms in different contexts of risk management, including optimal cash management [7] and inventory controls [16, 15, 39, 38]. More recent papers in the literature

[*] Received by the editors February 18, 2008; accepted for publication (in revised form) September 25, 2008; published electronically February 13, 2009.
http://www.siam.org/journals/sicon/48-2/71600.html

[†] Department of Industrial Engineering and Operations Research, University of California at Berkeley, Berkeley, CA 94720-1777 (xinguo@ieor.berkeley.edu).

[‡] Department of Mathematics, University of California at Berkeley, Berkeley, CA 94720-3840 (guoliang@math.berkeley.edu).

of mathematical finance include those on transaction cost in portfolio management [2, 21, 22, 10, 29, 32], insurance models [18, 5], liquidity risk [25, 4], optimal control of exchange rates [19, 30, 6], and finally, real options [40, 27].

Compared to regular controls, impulse control provides a more natural mathematical framework when the state space is discontinuous. It is a more general version of singular control allowing for nonzero fixed cost [15] and therefore harder to analyze. Indeed, in contrast to the singular/regular control theory, which enjoys a vast literature in financial engineering (see, for instance, Merton [28] and Karatzas and Shreve [20] among others), impulse control is less well understood, especially in terms of the structure of the optimal policy and regularity properties of the value function. In fact, regarding optimal policy, the best known work is perhaps still due to [7], which characterized the (u, U, d, D) form of the optimal policy for an inventory system. Although there have been various extensions of this structural result [16, 15, 39, 38, 34], most were derived through the verification theorem approach and by assuming a priori the smooth-fit property through the action/continuation regions. In the end, this approach usually amounts to solving complex algebraic equations that are hard to verify without a priori knowledge of the regularity property; thus the correctness of the "solution" is dubious. Indeed, there are (see, e.g., [41] and [3]) a few examples of singular control and stopping problems with explicit solutions, where value functions are not C^2, or even C^1, and only recently (see [26, 14], and [35]) the smooth-fit principle for one-dimensional singular control and the closely related switching control problems were established. In [1] value functions were shown to be the solutions of quasi-variational inequalities (QVIs) and the regularity properties were established for the case when the control is strictly positive and the state space is in a bounded region. However, to the best of our knowledge, regularity properties for value functions involving all-direction controls have not been fully established. This is an important omission in light of the wide range of applications mentioned earlier.

Our work. This paper studies regularity properties of the impulse control problem (1.3) on multidimensional diffusions in (2.1) subject to our conditions (A1)–(A4). Unlike the approach in [1], where the regularity was established through studying the corresponding QVIs, we first prove the value function to be the unique viscosity solution to the corresponding Hamilton–Jacobi–Bellman (HJB) equation. The main difficulties in proving the uniqueness of the viscosity solution are the unusual nonlocal property of the associated operator in the HJB equation and the unboundedness of the state space. We overcome these by exploiting and clarifying the definition of viscosity solutions in a local sense and by relating the problem to an optimal stopping problem (see also Remark 1). Next, we establish the regularity property of the value function, and in particular, the smooth-fit C^1 property through the boundaries between action and continuation regions. The existing technique in [1] does not apply here as it relies on a certain smoothness assumption that fails in our case (see also Remark 2). Finally, we show how to exploit this smooth-fit property to explicitly derive the form of optimal policy and the action/continuation regions for special cases that were first studied and analyzed in [7].

2. Formulation and assumptions.

2.1. Model formulation. Let us first define precisely the family of admissible controls. An *admissible impulse control* V consists of a sequence of stopping times $\tau_1, \tau_2, \ldots$ with respect to $\mathcal{F}_t$ (the natural filtration generated by W) and a corresponding sequence of $\mathbb{R}^n$-valued random variables $\xi_1, \xi_2, \dots$ satisfying the conditions

$$\begin{cases} 0 \le \tau_1 \le \tau_2 \le \cdots \le \tau_i \le \cdots, \\ \tau_i \to \infty \text{ almost surely as } i \to \infty, \end{cases}$$

and $\xi_i \in \mathcal{F}_{\tau_i}$ $\forall i \ge 1$.

As explained in the introduction, given an initial state $x \in \mathbb{R}^n$ and an admissible control $V = (\tau_1, \xi_1; \tau_2, \xi_2; \dots)$, the underlying process $X(t)$ is governed by the stochastic differential equation

$$\begin{cases} dX(t) = \mu(X(t-))dt + \sigma(X(t-))dW(t) + \sum_i \delta(t - \tau_i)\xi_i, \\ X(0) = x, \end{cases} \tag{2.1}$$

where $\delta(\cdot)$ denotes the Dirac delta function. Here the coefficients $\mu(\cdot)$ and $\sigma(\cdot)$ satisfy the Lipschitz conditions to ensure the existence and uniqueness of (1.1) (see, for instance, [37, Chapter V, Theorem 11.2]). Equation (2.1) is interpreted in a piecewise sense as in [1].

The associated total expected cost (objective function) is given by

$$J_x[V] := \mathbb{E}_x \left(\int_0^\infty e^{-rt} f(X(t))dt + \sum_{i=1}^\infty e^{-r\tau_i} B(\xi_i) \right), \tag{2.2}$$

where f is the "running cost," B is the "transaction cost," and $r > 0$ is the discount factor. We will specify the conditions on f and B in section 2.2 below.

The goal is to find the admissible $\widetilde{V}$ and the associated control sequence (τ_i, ξ_i) to minimize the total cost, i.e.,

$$J_x[\widetilde{V}] \le J_x[V] \quad \text{for any admissible } V.$$

We define the value function

$$u(x) = \inf_V J_x[V], \tag{2.3}$$

where the infimum is taken over all admissible control policies.

2.2. Assumptions and notations. Throughout this paper, we shall impose the following standing assumptions:

(A1) Lipschitz conditions on $\mu, \sigma : \mathbb{R}^n \to \mathbb{R}$: there exist constants $C_\mu, C_\sigma > 0$ such that

$$\begin{cases} |\mu(x) - \mu(y)| \le C_\mu |x - y| \\ |\sigma(x) - \sigma(y)| \le C_\sigma |x - y| \end{cases} \quad \forall x, y \in \mathbb{R}^n. \tag{2.4}$$

(A2) Lipschitz condition on the running cost $f \ge 0$: there exists a constant $C_f > 0$ such that

$$|f(x) - f(y)| \le C_f |x - y| \quad \forall x, y \in \mathbb{R}^n. \tag{2.5}$$

(A3) Conditions on the transaction cost function $B : \mathbb{R}^n \to \mathbb{R}$:

$$\begin{cases} \inf_{\xi \in \mathbb{R}^n} B(\xi) = K > 0, \\ B \in C(\mathbb{R}^n \backslash \{0\}), \\ |B(\xi)| \to \infty \text{ as } |\xi| \to \infty, \text{ and} \\ B(\xi_1) + B(\xi_2) \ge B(\xi_1 + \xi_2) + K \quad \forall \xi_1, \xi_2 \in \mathbb{R}^n. \end{cases} \tag{2.6}$$

(A4) $r > 2C_\mu + C_\sigma^2$.

We will also use the following notations for the operators:

$$\mathcal{M}\varphi(x) = \inf_{\xi\in\mathbb{R}^n} (\varphi(x+\xi) + B(\xi)), \tag{2.7}$$

$$\mathcal{L}\varphi(x) = -\operatorname{tr}\left[A \cdot D^2\varphi(x)\right] - \mu(x)\cdot D\varphi(x) + r\varphi(x), \tag{2.8}$$

where the matrix $A = (a_{ij})_{n\times n} = \frac{1}{2}\sigma(x)\sigma(x)^{\mathsf{T}}$.

Denote by $\Xi(x)$ the set of all the points ξ for which $\mathcal{M}u$ achieves the minimum value, where u is the value function, i.e.,

$$\Xi(x) := \{\xi \in \mathbb{R}^n : \mathcal{M}u(x) = u(x+\xi) + B(\xi)\}. \tag{2.9}$$

We also adopt the following standard notations for function spaces:

$$\begin{aligned}
UC(\mathbb{R}^n) &= \text{space of all uniformly continuous functions on } \mathbb{R}^n,\\
UC_{bb}(\mathbb{R}^n) &= \{f \in UC(\mathbb{R}^n) : f \text{ is bounded below}\},\\
W^{k,p}(U) &= \text{space of all } L^p \text{ functions with } \beta\text{th weak partial}\\
&\qquad \text{derivatives belonging to } L^p\ \forall|\beta| \le k,\\
C_c^\infty(U) &= \{f \in C^\infty(U) : f \text{ has compact support in } U\},\\
C^{k,\alpha}(D) &= \left\{ f \in C^k(D) : \sup_{x,y\in D} \left\{ \frac{|D^\beta f(x) - D^\beta f(y)|}{|x-y|^\alpha} \right\} < \infty \forall|\beta| \le k \right\}.
\end{aligned}$$

2.3. Preliminary results. We first establish some preliminary results about the value function, as well as the operator $\mathcal{M}$, under the standing assumptions.

LEMMA 2.1. *The value function $u(x)$ defined by* (2.3) *is Lipschitz.*

The proof of this lemma is a standard argument using Itô's formula and Gronwall's inequality. (See [23] and Theorem 10.1 in [12] for similar results and techniques.)

LEMMA 2.2 (basic properties of $\mathcal{M}$).

(1) *$\mathcal{M}$ is concave: for any $\varphi_1, \varphi_2 \in C(\mathbb{R}^n)$ and $0 \le \lambda \le 1$,*

$$\mathcal{M}(\lambda\varphi_1 + (1-\lambda)\varphi_2) \ge \lambda\mathcal{M}\varphi_1 + (1-\lambda)\mathcal{M}\varphi_2.$$

(2) *$\mathcal{M}$ is increasing: for any $\varphi_1 \le \varphi_2$ everywhere,*

$$\mathcal{M}\varphi_1 \le \mathcal{M}\varphi_2.$$

(3) *The operator $\mathcal{M}$ maps $C(\mathbb{R}^n)$ into $C(\mathbb{R}^n)$. In particular, $\mathcal{M}u(\cdot)$ is continuous. Moreover, $\mathcal{M}$ maps $UC(\mathbb{R}^n)$ into $UC(\mathbb{R}^n)$ and maps a Lipschitz function to a Lipschitz function.*

Proof. (1) and (2) are obvious.

(3) Suppose $\varphi \in C(\mathbb{R}^n)$. Then for any $x \in \mathbb{R}^n$, $\xi \in \mathbb{R}^n$, $\varepsilon > 0$,

$$-\varepsilon < \varphi(x+\xi+y) - \varphi(x+\xi) < \varepsilon,$$

provided that $|y| < \delta$ sufficiently small. Hence

$$\varphi(x+\xi) + B(\xi) - \varepsilon < \varphi(x+\xi+y) + B(\xi) < \varphi(x+\xi) + B(\xi) + \varepsilon.$$

This holds for arbitrary ξ, so by taking the infimum we get

$$\mathcal{M}\varphi(x) - \varepsilon \le \mathcal{M}\varphi(x+y) \le \mathcal{M}\varphi(x) + \varepsilon,$$

provided $|y| < \delta$ is small enough.

The last statement regarding Lipschitz functions can be proved similarly. □

LEMMA 2.3. *u and $\mathcal{M}u$ defined as above satisfy $u(x) \leq \mathcal{M}u(x) \,\forall\, x \in \mathbb{R}^n$.*

Proof. Suppose $x \in \mathbb{R}^n$, $\xi \in \mathbb{R}^n$, and $V = (\tau_1, \xi_1; \tau_2, \xi_2; \ldots)$ is an admissible control policy. Then $V' = (0, \xi; \tau_1, \xi_1; \tau_2, \xi_2; \ldots)$ is also admissible. Moreover,

$$u(x) \leq J_x[V'] = J_{x+\xi}[V] + B(\xi).$$

Taking the infimum over V and then the infimum over $\xi \in \mathbb{R}^n$, we get $u(x) \leq \mathcal{M}u(x)$. □

Now define the *continuation region* $\mathcal{C}$ and the *action region* $\mathcal{A}$ as follows:

$$\mathcal{C} := \{x \in \mathbb{R}^n : u(x) < \mathcal{M}u(x)\}, \tag{2.10}$$

$$\mathcal{A} := \{x \in \mathbb{R}^n : u(x) = \mathcal{M}u(x)\}. \tag{2.11}$$

Then, since u and $\mathcal{M}u$ are continuous, we have the following.

PROPOSITION 1. *$\mathcal{C}$ is open.*

PROPOSITION 2. *Suppose $x \in \mathcal{A}$; then*

(1) *the set*

$$\Xi(x) := \{\xi \in \mathbb{R}^n : \mathcal{M}u(x) = u(x+\xi) + B(\xi)\}$$

is nonempty; i.e., the infimum is in fact a minimum;

(2) *moreover, for any $\xi(x) \in \Xi(x)$, we have*

$$u(x + \xi(x)) \leq \mathcal{M}u(x + \xi(x)) - K,$$

in particular,

$$x + \xi(x) \in \mathcal{C}.$$

Proof. (1) Given $x \in \mathcal{A}$, take sequence $\{\xi_n\}$ such that

$$\mathcal{M}u(x) \leq u(x + \xi_n) + B(\xi_n) \leq \mathcal{M}u(x) + \frac{1}{n}.$$

Then $\{\xi_n\}$ is bounded since $|B(\xi)| \to \infty$ as $|\xi| \to \infty$. Extract a convergent subsequence $\{\xi_{n_k}\}$ that converges to ξ^*.

Claim. $\xi^* \neq 0$ and $\mathcal{M}u(x) = u(x+\xi^*) + B(\xi^*)$. Suppose $\xi^* = 0$. Since

$$u(x + \xi_{n_k}) + K \leq u(x + \xi_{n_k}) + B(\xi_{n_k}) \leq \mathcal{M}u(x) + \frac{1}{n_k},$$

by sending $k \to \infty$, we deduce that $u(x)+K \leq \mathcal{M}u(x) = u(x)$. This is a contradiction. Now that $\xi^* \neq 0$, clearly $\xi^* \in \Xi(x)$ since $B(\cdot)$ is also continuous at $\xi^* \neq 0$.

(2) Recall that $B(\xi_1) + B(\xi_2) \geq K + B(\xi_1 + \xi_2)$; hence

$$\begin{aligned}
\mathcal{M}u(x) &= \inf_{\xi \in \mathbb{R}} (u(x+\xi) + B(\xi)) \\
&= \inf_{\eta \in \mathbb{R}} (u(x + \xi(x) + \eta) + B(\xi(x) + \eta)) \\
&\leq \inf_{\eta \in \mathbb{R}} [u(x + \xi(x) + \eta) + B(\eta)] + B(\xi(x)) - K \\
&= \mathcal{M}u(x + \xi(x)) + B(\xi(x)) - K.
\end{aligned}$$

On the other hand, $\mathcal{M}u(x) = u(x + \xi(x)) + B(\xi(x))$. We get the desired result. □

Here we need $K > 0$ (assumption (A3)) to deduce $u(x+\xi(x)) < \mathcal{M}u(x + \xi(x))$.

3. Value function as viscosity solution. We show in this section that under certain conditions, the value function of the impulse control problem is the unique viscosity solution of the corresponding HJB equation

$$\text{(HJB)} \qquad \max(\mathcal{L}u - f, u - \mathcal{M}u) = 0.$$

3.1. Definition of viscosity solutions. First, recall (see [24]) the following definition of viscosity subsolutions (supersolutions, resp.):

(a) If $\varphi \in C^2(\mathbb{R}^n)$, $u - \varphi$ has a *global* maximum (minimum, resp.) at x_0 and $u(x_0) = \varphi(x_0)$, then

$$(3.1) \qquad \max(\mathcal{L}\varphi(x_0) - f(x_0), \varphi(x_0) - \mathcal{M}\varphi(x_0)) \le 0 \quad (\ge 0 \text{ resp.}).$$

However, note that the operator $\mathcal{M}$ is nonlocal; i.e., $\mathcal{M}\varphi(x_0)$ is not determined by values of φ in a neighborhood of x_0, and $\mathcal{M}\varphi(x_0)$ might be very small if φ is small away from x_0. Therefore, one has no control over $\mathcal{M}\varphi(x_0)$ by simply requiring that $u - \varphi$ have a local maximum (minimum, resp.) at x_0. In light of this, one can modify the definition of viscosity subsolutions (supersolutions, resp.) as follows: Suppose $u \in UC(\mathbb{R}^n)$.

(b) If $\varphi \in C^2(\mathbb{R}^n)$, $u - \varphi$ has a local maximum (minimum, resp.) at x_0 and $u(x_0) = \varphi(x_0)$, then

$$(3.2) \qquad \max(\mathcal{L}\varphi(x_0) - f(x_0), u(x_0) - \mathcal{M}u(x_0)) \le 0 \quad (\ge 0 \text{ resp.}).$$

In fact, one can show the following.

THEOREM 3.1. *The above two definitions of viscosity subsolutions (supersolutions, resp.) are equivalent.*

Proof. We will prove only the equivalence of subsolutions.

(b) ⇒ (a). Suppose $\varphi \in C^2(\mathbb{R}^n)$, $u - \varphi$ has a global maximum at x_0, and $u(x_0) = \varphi(x_0)$. Then $u \le \varphi$ globally, and by Lemma 2.2,

$$\varphi(x_0) = u(x_0) \le \mathcal{M}u(x_0) \le \mathcal{M}\varphi(x_0).$$

(a) ⇒ (b). Suppose $\varphi \in C^2(\mathbb{R}^n)$, $u - \varphi$ has a local maximum at x_0, and $u(x_0) = \varphi(x_0)$. For any $\varepsilon > 0$, take $r > 0$ so small that

$$u \le \varphi \le u + \varepsilon \text{ in } \bar{B}_{2r}(x_0) := \{x \in \mathbb{R}^n : |x - x_0| \le 2r\}.$$

There also exists a function $\tilde{\varphi} \in C^\infty(\mathbb{R}^n)$ such that

$$u \le \tilde{\varphi} \le u + \varepsilon \text{ in } \mathbb{R}^n.$$

(For instance, the usual mollification $\tilde{\varphi} = u * \eta^\delta + \varepsilon$, with $\delta > 0$ small enough.) Take a cutoff function $\zeta(x)$ such that

$$0 \le \zeta(x) \le 1;\ \zeta \equiv 1 \text{ on } \bar{B}_r(x_0);\ \zeta \equiv 0 \text{ off } \bar{B}_{2r}(x_0).$$

Now define

$$\psi(x) = \zeta(x)\varphi(x) + (1 - \zeta(x))\tilde{\varphi}(x).$$

Then clearly by construction,

$$u(x) \le \psi(x) \le u(x) + \varepsilon,$$

and ψ attains a global maximum at x_0. Thus

$$\mathcal{L}\psi(x_0) - f(x_0) \leq 0, \quad \psi(x_0) \leq \mathcal{M}\psi(x_0).$$

Note that by $\psi(x_0) = \varphi(x_0) = u(x_0)$, $D\psi(x_0) = D\varphi(x_0)$, $D^2\psi(x_0) = D^2\varphi(x_0)$, we have

$$\mathcal{L}\varphi(x_0) - f(x_0) \leq 0, \quad u(x_0) \leq \mathcal{M}\psi(x_0) \leq \mathcal{M}u(x_0) + \varepsilon.$$

Finally, since $\varepsilon > 0$ is arbitrary, by sending it to 0, we have (3.2). □

In light of Theorem 3.1, throughout the paper we shall adopt the following definition of viscosity solution.

DEFINITION 1. *The function u is called a viscosity solution of* (HJB) *if the following hold:*

(1) *(subsolution property.) For any $\varphi \in C^2(\mathbb{R}^n)$, if $u - \varphi$ has a local maximum at x_0 and $u(x_0) = \varphi(x_0)$, then we have*

$$\max(\mathcal{L}\varphi(x_0) - f(x_0), u(x_0) - \mathcal{M}u(x_0)) \leq 0. \tag{3.3}$$

(2) *(supersolution property.) For any $\varphi \in C^2(\mathbb{R}^n)$, if $u - \varphi$ has a local minimum at x_0 and $u(x_0) = \varphi(x_0)$, then we have*

$$\max(\mathcal{L}\varphi(x_0) - f(x_0), u(x_0) - \mathcal{M}u(x_0)) \geq 0. \tag{3.4}$$

Then we have the following known result [33]. (For the reader's convenience, we provide the proof in Appendix A.)

THEOREM 3.2. *The value function defined by* (2.3) *is a viscosity solution of the HJB equation*

$$\max\{\mathcal{L}u - f, u - \mathcal{M}u\} = 0. \tag{HJB}$$

3.2. Uniqueness of viscosity solution. In this section we shall show that the viscosity solution for (HJB) is unique in $UC_{bb}(\mathbb{R}^n)$.

The key idea is to relate the impulse control problem to an optimal stopping problem via the following operator $\mathcal{T}$, as in Bensoussan and Lions [1] and Ramaswamy and Dharmatti [36]. More precisely, given $\phi \in UC(\mathbb{R}^n)$, consider the following optimal stopping time problem:

$$\mathcal{T}\phi(x) := \inf_{\tau} \mathbb{E}\left(\int_0^\tau e^{-rt} f(X(t))dt + e^{-r\tau}\mathcal{M}\phi(x(\tau))\right) \tag{3.5}$$

subject to (1.1) and with the infimum taken over all $\mathcal{F}_t$ stopping times.

We shall first prove the uniqueness of the viscosity solution to the HJB equation (3.7) associated with this optimal stopping problem (3.5). We then exploit the properties of the operator $\mathcal{T}$ to establish the uniqueness of the viscosity solution to (HJB) for the impulse control problem.

3.2.1. Related optimal stopping problems. Now given (1.1), consider the following more generic optimal stopping problem with a terminal (nonnegative) cost $g(\cdot)$:

$$v(x) = \inf_{\tau} \mathbb{E}\left(\int_0^\tau e^{-rt} f(X(t))dt + e^{-r\tau} g(x(\tau))\right), \tag{3.6}$$

where f is the same as before, and the infimum is taken over all $\mathcal{F}_t$ stopping times.

First, the following result is well known [31].

PROPOSITION 3. *Assume that* $g \in C(\mathbb{R}^n)$. *Then the value function* $v(x)$ *defined by* (3.6) *is a continuous viscosity solution of the HJB equation*

$$\max\{\mathcal{L}w - f, w - g\} = 0 \text{ in } \mathbb{R}^n, \tag{3.7}$$

where $\mathcal{L}$ *is defined in* (2.8).

Next, we show the following.

THEOREM 3.3 (unique viscosity solution for optimal stopping). *Suppose* $g \in UC(\mathbb{R}^n)$ *and suppose there are some constants* $C, \Lambda > 0$ *such that*

$$\begin{cases} |\mu(x)| \leq C & \forall x \in \mathbb{R}^n, \\ a_{ij}(x)\xi_i\xi_j \leq \Lambda|\xi|^2 & \forall x, \xi \in \mathbb{R}^n, \end{cases} \tag{3.8}$$

where $(a_{ij}(x))_{n\times n} = \frac{1}{2}\sigma(x)\sigma(x)^{\mathsf{T}}$. *Then* (3.7) *has only one viscosity solution in* $UC(\mathbb{R}^n)$.

To prove Theorem 3.3, the following observation is useful.

LEMMA 3.4. *w is a viscosity solution of* $\max\{\mathcal{L}w - f,\, w - g\} = 0$ *if and only if it is a viscosity solution of*

$$F(x, w(x), Dw(x), D^2w(x)) = 0,^{1} \tag{3.9}$$

where $F : \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n \times S^n \to \mathbb{R}$ *is defined by*

$$F(x, t, p, X) = rt + \max\{G(x, p, X), -rg(x)\}, \tag{3.10}$$

$$G(x, p, X) = -\operatorname{tr}[AX] - \mu(x)\cdot p - f(x). \tag{3.11}$$

The key to proving Theorem 3.3 is the following comparison result.

LEMMA 3.5. *Let* $D \subset \mathbb{R}^n$ *be a bounded open set* $f, g \in UC(\mathbb{R}^n)$, *let* $v_* \in C(\overline{D})$ *be a subsolution, and let* $v^* \in C(\overline{D})$ *be a supersolution of*

$$\max\{\mathcal{L}w - f, w - g\} = 0 \text{ in } D.$$

Assume also that $v_* \leq v^*$ *on* ∂D. *Then* $v_* \leq v^*$ *in* $\overline{D}$.

The proof of Lemma 3.5 is based on the classical comparison theorem of second order degenerate elliptic differential equations in bounded domains (cf. [8]), and we defer it to Appendix B.

The following lemma (see Theorem 1 and the remarks on the fully nonlinear case in [9]) extends the above comparison result from bounded domains to an unbounded domain.

LEMMA 3.6. *Suppose* $F(x, t, p, X)$ *is elliptic in* X *such that there exist constants* $\alpha, C, \Lambda > 0$ *satisfying*

$$F(x, t + s, p + q, X + Y) \geq F(x, t, p, X) + \alpha s - C|q| - \Lambda\operatorname{tr}(Y) \tag{3.12}$$

$\forall x, p, q \in \mathbb{R}^n$, $t \in \mathbb{R}$, $s \geq 0$, $X, Y \in S^n$, $Y \geq 0$. *Suppose also that we have a comparison result for the equation* $F = 0$ *in bounded domains. If* v_* *and* v^* *are the*

[1] When there is no risk of confusion, we will also abbreviate (3.9) as $F = 0$.

continuous sub- and supersolution, respectively, of $F(x, w(x), Dw(x), D^2w(x)) = 0$ *in* $\mathbb{R}^n$ *with at most polynomial growth, then*

$$v_* \leq v^* \textit{ in } \mathbb{R}^n.$$

Now we can return to the following.

Proof of Theorem 3.3. By Lemmas 3.5 and 3.6, and noticing also

$$w \in UC(\mathbb{R}^n) \Rightarrow \sup_{x \in \mathbb{R}^n} \frac{|w(x)|}{1+|x|} < \infty,$$

it remains to show that F, defined by (3.10), satisfies (3.12) as follows:

$$\begin{aligned}
&F(x, t+s, p+q, X+Y) - F(x, t, p, X) \\
&\quad = rs + \max\{G(x, p+q, X+Y), -rg(x)\} - \max\{G(x, p, X), -rg(x)\} \\
&\quad \geq \begin{cases} rs \text{ if } G(x,p,X) \leq -rg(x), \\ rs + G(x, p+q, X+Y) - G(x,p,X) \text{ otherwise.} \end{cases}
\end{aligned}$$

If $Y \geq 0$, using (3.8), we have

$$\begin{aligned}
G(x, p+q, X+Y) - G(x,p,X) &= -\operatorname{tr}(AY) + \mu(x)q \\
&\geq -C|q| - \Lambda \operatorname{tr}(Y),
\end{aligned}$$

since $\operatorname{tr}(AY) = \operatorname{tr}(S^{\mathsf{T}}AS) \leq C \operatorname{tr}(S^{\mathsf{T}}S) = \Lambda \operatorname{tr}(Y)$, where $A = \frac{1}{2}\sigma\sigma^{\mathsf{T}}$, $Y = SS^{\mathsf{T}}$. This completes the proof. □

3.2.2. Uniqueness for impulse control problems. Now we are ready to prove the following.

THEOREM 3.7 (unique viscosity solution for impulse controls). *Assume that there are some constants* $C, \Lambda > 0$*, such that*

$$\begin{cases} |\mu(x)| \leq C & \forall x \in \mathbb{R}^n, \\ a_{ij}(x)\xi_i\xi_j \leq \Lambda|\xi|^2 & \forall x, \xi \in \mathbb{R}^n, \end{cases} \tag{3.8}$$

and $r > C_f$. *Then the value function for the impulse control problem defined by* (2.3) *is a unique viscosity solution in* $UC_{bb}(\mathbb{R}^n)$ *for the* HJB *equation.*

The proof relies on the following properties of the operator $\mathcal{T}$.

LEMMA 3.8.

(1) $\mathcal{T} : UC(\mathbb{R}^n) \to UC(\mathbb{R}^n)$.

(2) *If* $w \leq v$ *in* $\mathbb{R}^n$*, then* $\mathcal{T}w \leq \mathcal{T}v$ *in* $\mathbb{R}^n$.

(3) $\mathcal{T}$ *is concave on* $UC(\mathbb{R}^n)$.

Lemma 3.8 is immediate by the monotone and concave properties of $\mathcal{M}$ in Lemma 2.2, and by a direct application of Itô's formula and Gronwall's inequality.

Proof of Theorem 3.7. Suppose $w, v \in UC_{bb}(\mathbb{R}^n)$ are two solutions of (HJB). Without loss of generality, we can assume $w, v \geq 0$. Otherwise if $-C$ is a lower bound, then $w + C$, $v + C$ are nonnegative solutions to (HJB) of the same structure with f replaced by $f + rC$.

First, by definition of $\mathcal{T}$ together with Proposition 3 and Theorem 3.3, $\mathcal{T}w$ is the unique viscosity solution of

$$\max\{\mathcal{L}(\mathcal{T}w) - f, \mathcal{T}w - \mathcal{M}w\} = 0 \text{ in } \mathbb{R}^n.$$

On the other hand, w is also a viscosity solution of

$$\max\{\mathcal{L}w - f, w - \mathcal{M}w\} = 0 \text{ in } \mathbb{R}^n.$$

By uniqueness, $w = \mathcal{T}w$. Similarly, $v = \mathcal{T}v$.

Next, it suffices to show that if w, v satisfies $w - v \leq \gamma w$ for some $\gamma \in [0,1]$, then there exists some constant $\nu \in (0,1)$ such that

$$w - v \leq \gamma(1-\nu)w. \tag{3.13}$$

Indeed, if this claim holds, then note that $w - v \leq w$ since $v \geq 0$, and we have $w - v \leq (1-\nu)w$. Repeating the argument, we have

$$w - v \leq (1-\nu)^n w.$$

Sending $n \to \infty$, we get $w \leq v$. Switching the roles of w and v, we get $w = v$ as desired.

Now it remains to check the claim (3.13). First, by concavity of $\mathcal{T}$, we have

$$\mathcal{T}v \geq \mathcal{T}[(1-\gamma)w] \geq (1-\gamma)\mathcal{T}w + \gamma\mathcal{T}0.$$

Since f is at most with a linear growth, we deduce that

$$\begin{aligned} w = \mathcal{T}w &= \inf_{\tau} \mathbb{E}\left(\int_0^{\tau} e^{-rt} f(X(t))dt + e^{-r\tau}\mathcal{M}w(x(\tau))\right) \\ &\leq \mathbb{E}\int_0^{\infty} e^{-rt} f(X(t))dt \leq C\int_0^{\infty} e^{-rt}(1 + \mathbb{E}|X(t)|)dt \\ &\leq C(1+|x|)\int_0^{\infty} e^{-rt}e^{Ct}dt =: w_0 < \infty. \end{aligned}$$

Here, the last line follows from Gronwall's inequality.

Thus, $\mathcal{M}w(x) = \inf(w(x+\xi) + B(\xi)) \leq w_0 + K = K/\nu$, where

$$\nu := \frac{K}{K + w_0} \in (0,1).$$

Note that since $\mathcal{M}0 = K$, we obtain

$$\begin{aligned} \mathcal{T}0 &= \inf_{\tau} \mathbb{E}\left(\int_0^{\tau} e^{-rt} f(X(t))dt + e^{-r\tau}K\right) \\ &\geq \nu \inf_{\tau} \mathbb{E}\left(\int_0^{\tau} e^{-rt} f(X(t))dt + e^{-r\tau}\mathcal{M}w(X(t))\right) = \nu\mathcal{T}w. \end{aligned}$$

Therefore,

$$\mathcal{T}v \geq (1-\gamma)\mathcal{T}w + \gamma\mathcal{T}0 \geq (1-\gamma)\mathcal{T}w + \gamma\nu\mathcal{T}w.$$

The claim (3.13) is now clear by plugging in $\mathcal{T}w = w$, $\mathcal{T}v = v$. □

REMARK 1. *It is worth noting that in* [4] *the uniqueness of the viscosity solution was characterized for the value function of a finite horizon impulse control problem with execution delay involving the liquidity risk. However, the technique of* [4] *relies on the particular setup of a positive delay parameter and cannot be reduced to our case.*

4. Regularity of value function. We shall show in this section that the value function u is in the Sobolev space $W^{2,p}(\mathcal{O})$ for any open bounded region $\mathcal{O}$ and for any $p < \infty$, and in particular, $u \in C^1(\mathbb{R}^n)$. Throughout this section, we assume the operator $\mathcal{L}$ to be strictly elliptic: there exists a contant $c > 0$, such that $a_{ij}(x)\xi_i\xi_j \geq c|\xi|^2 \ \forall x, \xi \in R^n$.

Recall

$$\mathcal{C} := \{x \in \mathbb{R}^n : u(x) < \mathcal{M}u(x)\},$$
$$\mathcal{A} := \{x \in \mathbb{R}^n : u(x) = \mathcal{M}u(x)\}.$$

First, $u(\cdot)$ is clearly C^2 in the continuation region $\mathcal{C}$.

LEMMA 4.1 (C^2-regularity in $\mathcal{C}$). (1) *The value function $u(\cdot) \in C^2(\mathcal{C})$, and it satisfies the following differential equation in the classical sense:*

$$\mathcal{L}u(x) - f(x) = 0, \quad x \in \mathcal{C}. \tag{4.1}$$

(2) *For any set $D \Subset \mathcal{C}$,*[2]

$$u \in C^{2,\alpha}(\overline{D})$$

for some $\alpha > 0$.

Proof. Recall that from Theorem 3.2, u satisfies (4.1) in $\mathcal{C}$ in a viscosity sense. Now for any open ball $B \subset \mathcal{C}$, consider the Dirichlet problem

$$\begin{cases} \mathcal{L}w - f = 0 & \text{in } B, \\ w = u & \text{on } \partial B. \end{cases}$$

Classical Schauder estimates (cf. [13, Theorem 6.13]) ensure that such a solution w exists and belongs to $C^{2,\alpha}(B)$ because $f \in C^{0,\alpha}(B)$ for some $\alpha > 0$. Thus w satisfies the differential equation in a viscosity sense, whence $w = u$ in B by classical uniqueness results of the viscosity solution of a linear PDE in a bounded domain (cf. [8]). Hence,

$$u \in C^{2,\alpha}(B).$$

Finally, if $D \Subset \mathcal{C}$, then $\overline{D}$ can be covered by finitely many open balls contained in $\mathcal{C}$, and hence

$$u \in C^{2,\alpha}(\overline{D}). \qquad \square$$

Now we are are ready to establish the main regularity theorem.

THEOREM 4.2 ($W^{2,p}_{\text{loc}}$-regularity). *Assume that*

$$\sigma \in C^{1,1}(D) \textit{ for any compact set } D \subset \mathbb{R}^n. \tag{4.2}$$

Then for any bounded open set $\mathcal{O} \subset \mathbb{R}^n$, and $p < \infty$, we have

$$u \in W^{2,p}(\mathcal{O}).$$

In particular, $u \in C^1(\mathbb{R}^n)$, by the Sobolev embedding.

[2] $D \Subset \mathcal{C}$ means that D is *compactly contained* in $\mathcal{C}$; i.e., there exists a compact set F such that $D \subset F \subset \mathcal{C}$.

Proof. Given any bounded open set $\mathcal{O}$, we denote by $\mathcal{C}'$ ($\mathcal{A}'$, resp.) the restriction of the continuation (action, resp.) region within $\mathcal{O}$.

Our approach is to prove, for some constant C depending on $\mathcal{O}$,

$$-C \leq \mathcal{L}u \leq C, \tag{4.3}$$

in the sense of distribution. That is, for any smooth test function $\varphi \in C_c^\infty(\mathcal{O})$ with $\varphi \geq 0$, we have

$$-C\int_{\mathcal{O}} \varphi dx \leq \int_{\mathcal{O}} \left(a_{ij}u_{x_i}\varphi_{x_j} + b_i u_{x_i}\varphi + ru\varphi\right) dx \leq C\int_{\mathcal{O}} \varphi dx. \tag{4.4}$$

First, by (4.2), we can write the differential operator $\mathcal{L}$ in divergence form

$$\mathcal{L}u = -(a_{ij}u_{x_i})_{x_j} - b_i u_{x_i} + ru, \tag{4.5}$$

with $a_{ij} \in C^{1,1}(\overline{\mathcal{O}})$ and b_i Lipschitz. Note also that the first weak derivatives u_{x_i} are well defined since u is Lipschitz ($u \in W^{1,\infty}(\mathcal{O})$).

Observe that u is a viscosity subsolution of $\mathcal{L}u = f$ in $\mathcal{O}$ since it is a viscosity solution of (HJB). Thus u is also a distribution subsolution according to Ishii [17, Theorem 1]. Hence we have

$$\mathcal{L}u \leq C$$

in the sense of distribution, with $C = \sup_{\overline{\mathcal{O}}} |f|$.

Next, we show that

$$\mathcal{L}u^\varepsilon(x_0) \geq -C \quad \forall x_0 \in \mathcal{O}, \tag{4.6}$$

where $u^\varepsilon = u * \eta^\varepsilon \in C^\infty$ is the mollification of u, $\eta^\varepsilon = \eta(x/\varepsilon)/\varepsilon^n$, and $\eta(\cdot)$ is the standard mollifier. This is proved according to three different cases.

Case 1. $x_0 \in \mathcal{C}'$. Then by Lemma 4.1, $\mathcal{L}u(x_0) = f(x_0)$ in the classical sense. Hence

$$|\mathcal{L}u^\varepsilon(x_0)| = |f^\varepsilon(x_0)| \leq C,$$

where $C = \sup_{\overline{\mathcal{O}}} |f|$, which does not depend on x_0.

Case 2. $x_0 \in \partial\mathcal{A}'$. Then there exists a sequence $\{x_n\} \subset \mathcal{C}'$ converging to x_0. Since $|\mathcal{L}u^\varepsilon(x_n)| \leq C\ \forall n$ by the proof in Case 1, we obtain (4.6) by taking the limit as $n \to \infty$.

Case 3. $x_0 \in \operatorname{Int}\mathcal{A}'$, the interior of $\mathcal{A}'$. Since $\mathcal{A}' \subset \mathcal{O}$ is bounded, and $|B(\xi)| \to \infty$ as $|\xi| \to \infty$, we can find an open ball $\mathcal{O}' \supset \mathcal{O}$ so that $\xi(y) \in \mathcal{O}'\ \forall y \in \mathcal{A}'$ and $\xi(y) \in \Xi(y)$, because

$$B(\xi(y)) = \mathcal{M}u(y) - u(y + \xi(y)) \leq \mathcal{M}u(y) \leq \sup_{\overline{\mathcal{O}'}} \mathcal{M}u.$$

Now we define the set

$$D := \left\{ y \in \mathcal{O}' : u(y) < \mathcal{M}u(y) - \frac{K}{2} \right\}. \tag{4.7}$$

Clearly, $D \Subset \mathcal{C}$. From Lemma 4.1,

$$u \in C^{2,\alpha}(\overline{D}).$$

For any $x \in \operatorname{Int}\mathcal{A}'$, suppose $B_{\rho_1}(x) \subset \operatorname{Int}\mathcal{A}'$. Let us take $\xi(x) \in \Xi(x)$; then $y := x + \xi(x)$ satisfies

$$u(y) - \mathcal{M}u(y) \leq K$$

by Proposition 2. Hence $y \in D$.

On the other hand, since $u - \mathcal{M}u$ is uniformly continuous on $\overline{\mathcal{O}'}$, there exists $\rho_2 > 0$ such that

$$|y - y'| \leq \rho_2 \Rightarrow |u(y') - \mathcal{M}u(y') - (u(y) - \mathcal{M}u(y))| \leq \frac{K}{4}.$$

Therefore, for any constant $\lambda \in [-1, 1]$ and any unit vector $\chi \in \mathbb{R}^n$, $y' = y + \lambda\rho_2\chi$ satisfies

$$u(y') - \mathcal{M}u(y') \leq u(y) - \mathcal{M}u(y) + \frac{K}{4} < -\frac{K}{2}.$$

Hence, if we let $0 < \rho \leq \rho_1 \wedge \rho_2$, then $\forall\, \lambda \in [-1, 1]$, $\chi \in \mathbb{R}^n$, with $|\chi| = 1$, we have

$$\begin{aligned} &y = x + \xi(x) \in D, \quad y' = y + \lambda\rho\chi \in D,\\ &x \in \operatorname{Int}\mathcal{A}', \quad x + \lambda\rho\chi \in \operatorname{Int}\mathcal{A}'. \end{aligned}$$

By definition, we obtain

$$\begin{gathered} u(x) = \mathcal{M}u(x) = u(x + \xi(x)) + B(\xi(x)),\\ u(x \pm \rho\chi) = \mathcal{M}u(x \pm \rho\chi) \leq u(x \pm \rho\chi + \xi(x)) + B(\xi(x)), \end{gathered}$$

and hence the second difference quotient at x is

$$\begin{aligned} &\frac{1}{\rho^2}[u(x + \rho\chi) + u(x - \rho\chi) - 2u(x)]\\ &\quad \leq \frac{1}{\rho^2}[u(y + \rho\chi) + u(y - \rho\chi) - 2u(y)]\\ &\quad = \frac{1}{|\rho|}\int_0^1 [Du(y + \lambda\rho\chi) - Du(y - \lambda\rho\chi)] \cdot \chi d\lambda\\ &\quad \leq C_D, \end{aligned}$$

where $C_D = \sup_{x\in\overline{D}} |D^2u(x)| \leq \|u\|_{C^{2,\alpha}(\overline{D})}$.

Now, with $x_0 \in \operatorname{Int}\mathcal{A}'$ given, suppose $B_\theta(x_0) \subset \operatorname{Int}\mathcal{A}'$; then for any $0 < \varepsilon < \frac{\theta}{2}$, $\rho_1 = \frac{\theta}{2}$, $z \in B_\varepsilon(0)$, we have $B_{\rho_1}(x_0 - z) \subset \operatorname{Int}\mathcal{A}'$. Therefore, for $0 < \rho \leq \rho_1 \wedge \rho_2$ and unit vector $\chi \in \mathbb{R}^n$,

$$\begin{aligned} &\frac{1}{\rho^2}\left[u^\varepsilon(x_0 + \rho\chi) + u^\varepsilon(x_0 - \rho\chi) - 2u^\varepsilon(x_0)\right]\\ &= \frac{1}{\rho^2}\int_{B_\varepsilon(0)} [u(x_0 - z + \rho\chi) + u(x_0 - z - \rho\chi) - 2u(x_0 - z)]\eta^\varepsilon(z)\, dz\\ &\leq C_D \int_{B_\varepsilon(0)} \eta^\varepsilon(z)\, dz = C_D. \end{aligned}$$

Sending $\rho \to 0$ we get

$$\chi^\mathsf{T} D^2u^\varepsilon(x_0)\chi \leq C_D.$$

Hence,

$$\begin{aligned}\operatorname{tr}(\sigma(x_0)\sigma(x_0)^{\mathsf{T}}D^2u^\varepsilon(x_0)) &= \operatorname{tr}(\sigma^{\mathsf{T}}(x_0)D^2u^\varepsilon(x_0)\sigma(x_0))\\ &= \sum_k \sigma_k^{\mathsf{T}} D^2u^\varepsilon \sigma_k\\ &\le C_D\sum_{i,j}|\sigma_{ij}(x_0)|^2\\ &\le C,\end{aligned}$$

where σ_k is the kth column of the matrix σ, σ_{ij} is the (i,j)th element of σ, and the last inequality is due to continuity of σ.

Note that $|u^\varepsilon(x_0)| + |Du^\varepsilon(x_0)| \le \|u\|_{W^{1,\infty}(\overline{\mathcal{O}})}$ and $\mu(x)$ bounded; we deduce

$$\mathcal{L}u^\varepsilon(x_0) = -\frac{1}{2}\operatorname{tr}\left(\sigma(x_0)\sigma(x_0)^{\mathsf{T}}D^2u^\varepsilon(x_0)\right) - \mu(x_0)\cdot Du^\varepsilon(x_0) + ru^\varepsilon(x_0) \ge -C,$$

where C is independent of x_0.

Finally, (4.6) implies that for any smooth test function $\varphi \in C_c^\infty(\mathcal{O})$, $\varphi \ge 0$,

$$\int_{\mathcal{O}} \left(a_{ij}u^\varepsilon_{x_i}\varphi_{x_j} + b_i u^\varepsilon_{x_i}\varphi + ru^\varepsilon\varphi\right)dx \ge -C\int_{\mathcal{O}}\varphi dx. \tag{4.8}$$

Since $u \in W^{1,2}(\mathcal{O})$, $u^\varepsilon \to u$ in $L^2(\mathcal{O})$, and $u^\varepsilon_{x_i} \to u_{x_i}$ in $L^2(\mathcal{O})$. Sending $\varepsilon \to 0$ in the above inequality, we obtain (4.3). Therefore,

$$\mathcal{L}u \in L^\infty(\mathcal{O}).$$

By the Calderón–Zygmund estimate (see, e.g., [13]),

$$u \in W^{2,p}(\mathcal{O}) \quad \forall p < \infty. \qquad \square$$

Remark 2. *Compared to the regularity results of Bensoussan and Lions* [1], *we deal with a control which is unbounded and not necessarily positive. Moreover, we prove the regularity property for the value function as a viscosity solution of the HJB equation as opposed to their weak solutions of QVIs with $u \in H_0^1$ satisfying*

$$\begin{cases} a(u, v-u) \ge \langle f, v-u\rangle & \forall v \in H_0^1,\ v \le \mathcal{M}u,\\ 0 \le u \le \mathcal{M}u, \end{cases}$$

where $a(\phi,\psi) = \langle \mathcal{L}\phi, \psi\rangle$ and $\langle\cdot,\cdot\rangle$ is the paring between the Hilbert space H_0^1 and its dual space. (See Lemmas 2.3–2.4 *and Theorem* 2.2 *of Chapter* 4 *in* [1]*). In addition, their key lemma, Lemma* 2.3*, requires, in our notation, $\gamma(x) := \inf_{x+\xi\in\partial O} B(\xi) \in W^{2,\infty}(O)$, and therefore C^1. However, this condition fails in our case; see our example in section* 5*, where the corresponding γ has a corner.*

5. Structure of the value function. Having obtained the regularity results for the value function, in this section we shall characterize the structure of the value function as well as the continuation/action regions for the following special case: $n = 1$ and the cost functions f and B are given by

$$f(x) = \begin{cases} hx & \text{if } x \ge 0,\\ -px & \text{if } x \le 0, \end{cases} \tag{5.1}$$

$$B(\xi) = \begin{cases} K^+ + k^+\xi & \text{if } \xi \ge 0,\\ K^- - k^-\xi & \text{if } \xi < 0, \end{cases} \tag{5.2}$$

where h, p, k^+, k^-, K^+, K^- are positive constants. Moreover, we assume that μ and σ are all constant:

$$\sigma(x) \equiv \sigma, \quad \mu(x) \equiv \mu. \tag{5.3}$$

In addition, we will impose the following condition to rule out triviality:

$$h - rk^- > 0, \quad p - rk^+ > 0. \tag{5.4}$$

This case was first characterized in [7] with a verification-type argument. Here we provide an alternative derivation by exploiting the regularity property established earlier. We shall show directly the following.

THEOREM 5.1 (characterization of the solution structure). *Assuming* (5.1), (5.2), (5.3), *and* (5.4),

(1) *there exist constants* $-\infty < q < s < \infty$ *such that*

$$\begin{aligned} \mathcal{C} &:= \{x \in \mathbb{R} : u(x) < \mathcal{M}u(x)\} = (q, s), \\ \mathcal{A} &:= \{x \in \mathbb{R} : u(x) = \mathcal{M}u(x)\} = (-\infty, q] \cup [s, +\infty); \end{aligned}$$

(2) *the value function* u *defined in* (2.3) *satisfies*

$$\begin{cases} \mathcal{L}u(x) = f(x), & q < x < s, \\ u(x) = u(s) + k^-(x - s), & x \geq s, \\ u(x) = u(q) + (q - x)k^+, & x \leq q; \end{cases}$$

(3) *there are points* $Q, S \in (q, s)$ *such that*

$$\begin{aligned} u'(q) = u'(Q) &= -k^+, & u(q) &= u(Q) + K^+ + k^+(Q - q), \\ u'(s) = u'(S) &= k^-, & u(s) &= u(S) + K^- - k^-(S - s), \end{aligned}$$

as shown in Figure 5.1.

The proof of Theorem 5.1 is based on a series of lemmas.

LEMMA 5.2. *Under assumptions* (5.1) *and* (5.2), *for any* $x_0 \in \mathcal{A}$ *and*

$$\xi_0 \in \Xi(x_0) := \{\xi \in \mathbb{R} : \mathcal{M}u(x_0) = u(x_0 + \xi) + B(\xi)\},$$

we have

$$u'(x_0) = u'(x_0 + \xi_0) = \begin{cases} -k^+, & \xi_0 > 0, \\ k^-, & \xi_0 < 0. \end{cases} \tag{5.5}$$

Proof. First, such a ξ_0 exists and $\xi_0 \neq 0$ by Proposition 2. By definition, $u(x_0) = \mathcal{M}u(x_0) = u(x_0 + \xi_0) + B(\xi_0)$, which means ξ_0 is a global minimum of the function $u(x_0 + \cdot) + B(\cdot)$. Also, $\xi_0 \neq 0$ implies that B is also differentiable at ξ_0, and hence

$$u'(x_0 + \xi_0) = -B'(\xi_0) = \begin{cases} -k^+, & \xi_0 > 0, \\ k^-, & \xi_0 < 0. \end{cases}$$

Now, for any $\delta \neq 0$, we have

$$u(x_0 + \delta) \leq \mathcal{M}u(x_0 + \delta) \leq u(x_0 + \delta + \xi_0) + B(\xi_0).$$

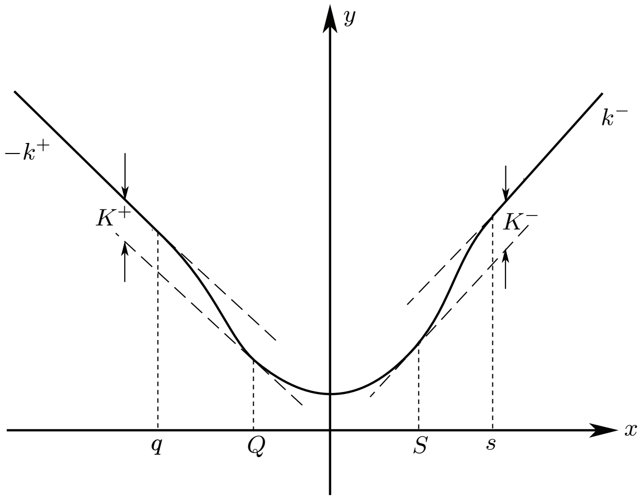


FIG. 5.1. *The value function u.*

Thus,

$$\frac{u(x_0+\delta)-u(x_0)}{\delta} \le \frac{u(x_0+\xi_0+\delta)-u(x_0+\xi_0)}{\delta}, \quad \delta > 0,$$
$$\frac{u(x_0+\delta)-u(x_0)}{\delta} \ge \frac{u(x_0+\xi_0+\delta)-u(x_0+\xi_0)}{\delta}, \quad \delta < 0.$$

Taking the limit as $\delta \to 0^+$ ($\delta \to 0^-$, resp.), we conclude that

$$u'(x_0) = u'(x_0+\xi_0). \qquad \square$$

LEMMA 5.3. *Assume* (5.1), (5.2), *and* (5.3). *For any* $x_0 \in \mathcal{A}$ *and* $\xi_0 \in \Xi(x_0)$,
(1) *if* $x_0 > 0$, *then* $\xi_0 < 0$ *and* $u'(x_0) = k^-$;
(2) *if* $x_0 < 0$, *then* $\xi_0 > 0$ *and* $u'(x_0) = -k^+$.

Proof. (1) Suppose not; then there exists a $\xi_0 \in \Xi(x_0)$ with $\xi_0 > 0$ and

$$u(x_0) = \mathcal{M}u(x_0) = u(x_0+\xi_0) + B(\xi_0).$$

First, take an ε-optimal strategy $V = (\tau_1, \xi_1; \tau_2, \xi_2; \ldots)$ for the initial level $x_0+\xi_0$, i.e.,

$$J_{x_0+\xi_0}[V] \le u(x_0+\xi_0) + \varepsilon,$$

where $\varepsilon > 0$ is arbitrarily small, to be chosen later.

Construct a strategy for x_0,

$$V_1 = (0, \xi_0; \tau_1, \xi_1; \tau_2, \xi_2; \ldots).$$

Then by definition,

$$J_{x_0}[V_1] = J_{x_0+\xi_0}[V] + B(\xi_0) \le u(x_0+\xi_0) + \varepsilon + B(\xi_0) = u(x_0) + \varepsilon. \tag{5.6}$$

On the other hand, we can construct another strategy for x_0,

$$V_2 = (\tau, \xi_0; \tau_1, \xi_1; \tau_2, \xi_2; \ldots),$$

where

$$\tau = \inf\{t : x(t; x_0, V) < 0\}.$$

Here, we use $x(t; x_0, V)$ to denote the solution of (2.1) with initial value x_0 and strategy V.

Since the system is linear by (5.3), we have

$$x(t; x_0, V_2) = \begin{cases} x(t; x_0, V_1) - \xi_0 > 0, & t < \tau; \\ x(t; x_0, V_1), & t \geq \tau. \end{cases}$$

It follows that $f(x(t; x_0, V_2)) \leq f(x(t; x_0, V_1)) \,\forall t$. So

$$\begin{aligned} &J_{x_0}[V_2] - J_{x_0}[V_1] \\ &\quad = \mathbb{E}\left(\int_0^\infty [f(x(t; x_0, V_2)) - f(x(t; x_0, V_1))] e^{-rt} dt + e^{-r\tau} B(\xi_0) - B(\xi_0) \right) \\ &\quad \leq B(\xi_0)(\mathbb{E} e^{-r\tau} - 1) =: -\nu. \end{aligned} \tag{5.7}$$

It remains to show that $\nu > 0$.

Claim. $\tau > 0$ a.s. if ε is sufficiently small.

Clearly, $\tau \geq \tau_1 \wedge \inf\{t : x_0 + \mu t + \sigma W(t) < 0\}$ and obviously $\inf\{t : x_0 + \mu t + \sigma W(t) < 0\} > 0$ a.s., since $x_0 > 0$. Now we need to prove $\tau > 0$ a.s. Suppose not; then

$$\begin{aligned} J_{x_0+\xi_0}[V] &= \mathbb{E}\left\{ J_{x_0+\xi_0+\xi_1}[V \setminus (\tau_1, \xi_1)] + B(\xi_1) \right\} \\ &\geq \mathbb{E}\{u(x_0 + \xi_0 + \xi_1) + B(\xi_1)\} \\ &\geq \mathcal{M}u(x_0 + \xi_0). \end{aligned}$$

However, since $x_0 + \xi_0 \in \mathcal{C}$ by Proposition 2, if we take $0 < \varepsilon \leq (\mathcal{M}u(x_0 + \xi_0) - u(x_0 + \xi_0))/2$, we have

$$J_{x_0+\xi_0}[V] \leq u(x_0 + \xi_0) + \varepsilon = \mathcal{M}u(x_0 + \xi_0) - \varepsilon.$$

This is a contradiction. Thus we proved the claim, and it follows that $\nu = -B(\xi_0)(\mathbb{E}e^{-r\tau} - 1) > 0$.

Combining (5.6) and (5.7) and taking $\varepsilon < \nu/2$,

$$J_{x_0}[V_2] \leq J_{x_0}[V_1] - \nu \leq u(x_0) + \varepsilon - \nu < u(x_0) - \nu/2.$$

This is a contradiction, and we have $\xi_0 > 0$. It follows from Lemma 5.2 that $u'(x_0) = k^-$.

The proof of (2) is exactly the same. □

Recall $\mathcal{C}$ is an open set, i.e., a union of open intervals. The following lemma rules out the possibility that $\mathcal{C}$ contains unbounded intervals.

LEMMA 5.4. *Under assumption* (5.4), $\mathcal{C}$ *does not contain any of the intervals* $(a, +\infty)$ *or* $(-\infty, b)$, *with* $a \geq -\infty$, $b \leq +\infty$.

Proof. Suppose $\mathcal{C} \supset (a, +\infty)$. Then we have, for $c > \max\{a, 0\}$,

$$-\frac{1}{2}\sigma^2 u'' - \mu u' + ru = hx, \quad x \in (c, +\infty).$$

The ODE has a general solution

$$u(x) = C_1 e^{\lambda_1 x} + C_2 e^{\lambda_2 x} + \frac{hx}{r} + \frac{\mu h}{r^2},$$

where $\lambda_1 = \frac{-\mu - \sqrt{\mu^2 + 2\sigma^2 r}}{\sigma^2} < 0$, $\lambda_2 = \frac{-\mu + \sqrt{\mu^2 + 2\sigma^2 r}}{\sigma^2} > 0$. Note that $C_2 = 0$; otherwise $u'(x) = C_1 \lambda_1 e^{\lambda_1 x} + C_2 \lambda_2 e^{\lambda_2 x} + \frac{h}{r}$ is unbounded, approaching $+\infty$ or $-\infty$ as $x \to +\infty$, which contradicts the fact that u is Lipschitz.

Now, for any $x > c > 0$,

$$C_1 e^{\lambda_1 x} + \frac{hx}{r} + \frac{\mu h}{r^2} = u(x) < \mathcal{M}u(x) \le u(c) + K^- - k^-(c - x)$$

or

$$\left(\frac{h}{r} - k^-\right) x + C_1 e^{\lambda_1 x} < u(c) + K^- - k^- c - \frac{\mu h}{r^2}.$$

As $x \to +\infty$, we get a contradiction, noticing that $h - rk^- > 0$.

$p - rk^+ > 0$ will ensure that $\mathcal{C}$ cannot contain intervals of $(-\infty, b)$ type. Therefore, we prove the lemma. □

Finally, we see the following.

LEMMA 5.5. *Assume* (5.1), (5.2), (5.3), *and* (5.4). *Then* $\mathcal{C}$ *is connected.*

Proof. Suppose not. We prove by contradiction through the following steps.

Step 1. By assumption, there are points $y_1 < y_2 < y_3$ so that $y_1, y_3 \in \mathcal{C}$ while $y_2 \in \mathcal{A}$. Define

$$x_1 := \inf\{x \in \mathcal{A} : x \le y_2, [x, y_2] \subset \mathcal{A}\},$$
$$x_2 := \sup\{x \in \mathcal{A} : x \ge y_2, [y_2, x] \subset \mathcal{A}\}.$$

Clearly, x_1, x_2 exist and are finite, with $[x_1, x_2] \subset \mathcal{A}$. (We do not rule out the possibility that $x_1 = x_2$ here.)

By Lemma 5.2, $u'(x) = k^-$ or $-k^+$, for any $x \in \mathcal{A}$. Since $u \in C^1(\mathbb{R})$, u' is a constant on $[x_1, x_2]$. Assume $u'(x) = k^- \, \forall x \in [x_1, x_2]$, and consider u at the point x_2. (The other case $u' = -k^+$ is similar. In that case we consider the point x_1 instead.)

Step 2. We show that

$$u(x) \le u(x_2) + k^-(x - x_2) \quad \forall x \ge x_2, \tag{5.8}$$

and the inequality is strict if $x > x_2$ and $x \in \mathcal{C}$.

Let $\xi_2 \in \Xi(x_2)$. Then $\xi_2 < 0$ by Lemma 5.3, and hence $B(\xi_2) = K^- - k^- \xi_2 = B(\xi_2 - y) - k^- y$ for $y = x - x_2 \ge 0$. Therefore,

$$\begin{aligned} u(x) &\le (\text{or} < \text{if } x \in \mathcal{C})\ \mathcal{M}u(x) \le u(x_2 + \xi_2) + B(x_2 + \xi_2 - x) \\ &= u(x_2 + \xi_2) + B(\xi_2) + k^-(x - x_2) = u(x_2) + k^-(x - x_2). \end{aligned}$$

Step 3. We show that

$$-\mu k^- + ru(x_2) \le hx_2. \tag{5.9}$$

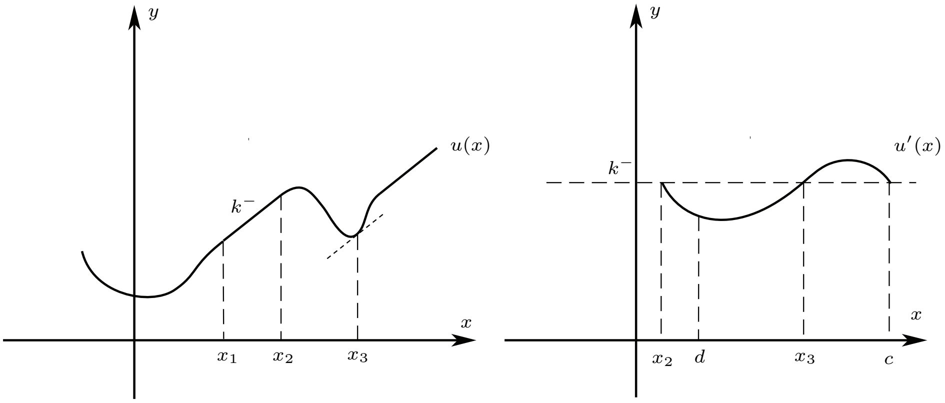


FIG. 5.2. *Proof of Lemma* 5.5.

Since $x_2 \in \mathcal{A}$, $u'(x_2) = k^-$, Lemma 5.3 implies that $x_2 \geq 0$. However, (5.8) implies that x_2 is a local maximum of $u - \phi$, where $\phi(x) = u(x_2) + k^-(x - x_2)$ is linear. By the viscosity subsolution property, we have $\mathcal{L}\phi(x_2) \leq f(x_2) = hx_2$, which is (5.9).

Step 4. There exists a point $x_3 > x_2$ such that

$$-\mu k^- + ru(x_3) \geq hx_3. \tag{5.10}$$

Suppose (x_2, c) is an open interval component of $\mathcal{C}$. Then by Lemma 5.4, $c < \infty$ and thus $c \in \mathcal{A}$. Lemma 5.3 implies that $u'(c) = k^-$ since $c > x_2 \geq 0$. Take $d \in (x_2, c)$ such that $u'(d) < k^-$. Such a d exists since $u(x) < u(x_2) + k^-(x - x_2)$ for $x \in (x_2, c)$. (See Figure 5.2.) Let

$$x_3 = \inf\{d \leq x \leq c : u'(x) = k^-\},$$

which is well defined, since c is in this set. Clearly $x_3 > d > x_2$. Moreover, $u'(x) < k^- = u'(x_3)$ for $d \leq x < x_3$ by definition. So

$$u''(x_3) \geq 0.$$

Thus $hx_3 = -\frac{1}{2}\sigma^2 u''(x_3) - \mu u'(x_3) + ru(x_3) \leq -\mu k^- + ru(x_3)$.

Step 5. From (5.9) and (5.10), it follows that

$$u(x_3) - u(x_2) \geq h/r(x_3 - x_2) > k^-(x_3 - x_2),$$

by (5.4). This is a contradiction to (5.8). □

Proof of Theorem 5.1. (1) Since $\mathcal{C}$ is connected, by Lemma 5.4, $\mathcal{C} = (q, s)$ for some $-\infty < q < s < \infty$.

(2) Suppose $x \geq s$ and $\xi \in \Xi(x)$. Because $x + \xi \in \mathcal{C} = (q, s)$, we have $\xi < 0$, whence $u'(x) = k^-$ by Lemma 5.2. Thus, $u(x) = u(s) + k^-(x - s)$ for $x \geq s$. A similar argument shows that $u(x) = u(q) + (q - x)k^+$, for $x \leq q$.

(3) Let $\xi \in \Xi(s)$ and $S = s + \xi$. Then $S \in (q, s)$, $u'(s) = u'(S) = k^-$, and

$$u(s) = \mathcal{M}u(s) = u(S) + B(S - s) = u(S) + K^- - k^-(S - s).$$

The remaining statement is similar. □

Appendix.

Appendix A. Proof of Theorem 3.2.

Proof. (1) (subsolution property.) Suppose $\varphi \in C^2(\mathbb{R}^n), u - \varphi$ has a local maximum at x_0, and $u(x_0) = \varphi(x_0)$. By Lemma 2.3, it suffices to prove

$$\mathcal{L}\varphi(x_0) - f(x_0) \le 0.$$

For any admissible control $V = \{\tau_1, \xi_1; \tau_2, \xi_2; \ldots\}$ and $\tau > 0$, the control $V' = \{\tau_1 + \tau, \xi_1; \tau_2 + \tau, \xi_2; \ldots\}$ is also admissible, and thus

$$u(x_0) \le J_{x_0}[V'] = \mathbb{E}\left(\int_0^\tau f(X(t))e^{-rt}dt + e^{-r\tau}J_{x(\tau)}[V]\right),$$

which implies

$$u(x_0) \le \mathbb{E}\left(\int_0^\tau f(X(t))e^{-rt}dt + e^{-r\tau}u(x(\tau))\right), \tag{A.1}$$

where $X(t)$ is the solution of

$$\begin{cases} dX(t) = \mu(X(t))dt + \sigma(X(t))dW(t), \\ X(0) = x_0. \end{cases} \tag{A.2}$$

Meanwhile, Dynkin's formula gives

$$\mathbb{E}\left(e^{-r\tau}\varphi(x(\tau))\right) - \varphi(x_0) = -\mathbb{E}\left(\int_0^\tau e^{-rt}\mathcal{L}\varphi(X(t))dt\right). \tag{A.3}$$

Noting that $u \le \varphi$ near x_0 and $u(x_0) = \varphi(x_0)$, combining (A.1) and (A.3), and sending $\tau \to 0^+$, we have $\mathcal{L}\varphi - f \le 0$ at x_0.

(2) (supersolution property.) Suppose $\varphi \in C^2(\mathbb{R}^n)$, $u - \varphi$ has a local minimum at x_0, and $u(x_0) = \varphi(x_0)$. If $u(x_0) = \mathcal{M}u(x_0)$, then (3.4) is trivially true. Thus we assume $u(x_0) < \mathcal{M}u(x_0)$. By continuity of $\mathcal{M}$, there exist constants $\delta > 0, \rho > 0$ such that

$$\varphi(x) \le u(x), u(x) - \mathcal{M}u(x) < -\delta \text{ whenever } |x - x_0| < \rho. \tag{A.4}$$

Define

$$\tau_\rho := \inf\{t > 0 : |X(t) - x_0| \ge \rho\}.$$

For any $\varepsilon > 0$, choose an ε-optimal control $V = (\tau_1, \xi_1; \tau_2, \xi_2; \ldots)$, i.e.,

$$J_{x_0}[V] \le u(x_0) + \varepsilon.$$

Then for any stopping time $\tau \le \tau_1$ a.s.,

$$u(x_0) + \varepsilon \ge J_{x_0}[V] = \mathbb{E}\left(\int_0^\tau f(X(t))e^{-rt}dt + e^{-r\tau}J_{x(\tau^-)}[V']\right), \tag{A.5}$$

where $V' = (\tau_1 - \tau, \xi_1; \tau_2 - \tau, \xi_2; \ldots)$ is admissible.

Fix $R > 0$ and let $\bar{\tau} = \tau_\rho \wedge R$.

Claim. $\mathbb{P}\{\tau_1 < \bar{\tau}\} \to 0$ as $\varepsilon \to 0$.

Consider (A.5) with $\tau = \tau_1$. On the set $\{\tau_1 < \bar{\tau}\}$,

$$\begin{aligned} J_{x(\tau_1^-)}[V'] &= \mathbb{E}\left(J_{x(\tau_1^-)+\xi_1}[\tilde{V}] + B(\xi_1)\right) \\ &\geq \mathbb{E}\left(u(x(\tau_1^-) + \xi_1) + B(\xi_1)\right) \geq \mathcal{M}u(x(\tau_1^-)) \\ &\geq u(x(\tau_1^-)) + \delta, \end{aligned}$$

because of (A.4). Otherwise, we still have $J_{x(\tau_1^-)}[V'] \geq u(x(\tau_1^-))$. Thus,

$$\begin{aligned} u(x_0) + \varepsilon &\geq \mathbb{E}\left(\int_0^{\tau_1} f(X(t))e^{-rt}dt + e^{-r\tau_1} J_{x(\tau_1^-)}[V''] \right) \\ &\geq \mathbb{E}\left(\int_0^{\tau_1} f(X(t))e^{-rt}dt + e^{-r\tau_1} u(x(\tau_1^-)) \right) + e^{-rR}\delta \cdot \mathbb{P}\{\tau_1 < \bar{\tau}\} \\ &\geq u(x_0) + e^{-rR}\delta \cdot \mathbb{P}\{\tau_1 < \bar{\tau}\}. \end{aligned}$$

This proves the claim.

Take $\tau = \bar{\tau} \wedge \tau_1$; by sending $\varepsilon \to 0$ in (A.5), we get

$$u(x_0) \geq \mathbb{E}\left(\int_0^{\bar{\tau}} f(X(t))e^{-rt}dt + e^{-r\bar{\tau}} u(x(\bar{\tau}^-)) \right).$$

Note that $\varphi(x(\bar{\tau}^-)) \leq u(x(\bar{\tau}^-))$ and $\varphi(x_0) = u(x_0)$, and that the above inequality together with Dynkin's formula (A.3) gives

$$\mathbb{E}\left(\int_0^{\bar{\tau}} e^{-rt}(\mathcal{L}\varphi - f)(X(t))dt \right) \geq 0.$$

Dividing by $\mathbb{E}(\bar{\tau})$ and sending $\rho \to 0$, we obtain the desired result (3.4). □

Appendix B. Proof of Lemma 3.5. To prove this lemma, we first recall a well-known comparison theorem on elliptic PDEs.

Theorem B.1 (Theorem 3.3 in [8]). *Let U be a bounded open subset of $\mathbb{R}^n$, and let $F \in C(U \times \mathbb{R} \times \mathbb{R}^n \times S^n)$ satisfy the following:*

(1) *$F(x, t, p, X) \leq F(x, s, p, Y)$ whenever $t \leq s$, $Y \leq X$.*

(2) *There exists some $\gamma > 0$ such that, for $r \geq s$ and $(x, p, X) \in \bar{U} \times \mathbb{R}^n \times S^n$,*

$$\gamma(r - s) \leq F(x, r, p, X) - F(x, s, p, X).$$

(3) *There is a function $\omega : [0, \infty] \to [0, \infty]$ with $\omega(0+) = 0$ such that*

$$F(y, r, \alpha(x - y), Y) - F(x, r, \alpha(x - y), X) \leq \omega(\alpha|x - y|^2 + |x - y|)$$

whenever $x, y \in U$, $r \in \mathbb{R}$, $X, Y \in S^n$, and

$$-3\alpha \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} \leq \begin{pmatrix} X & 0 \\ 0 & Y \end{pmatrix} \leq 3\alpha \begin{pmatrix} I & -I \\ -I & I \end{pmatrix}.$$

Let $u \in C(\bar{U})$ be a viscosity subsolution, and let $v \in C(\bar{U})$ be a viscosity supersolution of $F = 0$ in U with $u \leq v$ on ∂U. Then $u \leq v$ in $\bar{U}$.

Here S^n is the collection of $n \times n$ real symmetric matrices equipped with the usual ordering and I is the identity matrix.

Proof of Lemma 3.5. In view of Lemmas 3.4 and 3.5, it suffices to verify that F, defined by (3.10), satisfies the conditions of Theorem B.1. Clearly, F is continuous, and

$$F(x,t,p,X) \le F(x,s,p,Y) \text{ whenever } t \le s, Y \le X.$$

$$F(x,t,p,X) - F(x,s,p,X) = r(t-s).$$

Finally, we are to prove that there exists a function $\omega : [0,\infty] \to [0,\infty]$ satisfying $\omega(0+) = 0$ such that if $x, y \in D$, $t \in \mathbb{R}$, and X, Y are symmetric and satisfying for some $\alpha > 0$,

$$-3\alpha \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} \le \begin{pmatrix} X & 0 \\ 0 & -Y \end{pmatrix} \le 3\alpha \begin{pmatrix} I & -I \\ -I & I \end{pmatrix},$$

then

$$F(y,t,\alpha(x-y),Y) - F(x,t,\alpha(x-y),X) \le \omega(\alpha|x-y|^2 + |x-y|). \tag{B.1}$$

It is easy to check that $G(x,p,X)$ and $-rg(x)$ satisfy (B.1), since $f, g \in UC(\mathbb{R}^n)$ (cf. Example 3.6 in [8]). Hence

$$\begin{aligned} &F(y,t,\alpha(x-y),Y) - F(x,t,\alpha(x-y),X) \\ &= \max\{G(y,\alpha(x-y),Y), -rg(y)\} - \max\{G(x,\alpha(x-y),X), -rg(x)\} \\ &\le \begin{cases} -r(g(y)-g(x)) & \text{if } G(y,\alpha(x-y),Y) \le -rg(y), \\ G(y,\alpha(x-y),Y) - G(x,\alpha(x-y),X) & \text{otherwise} \end{cases} \\ &\le \omega(\alpha|x-y|^2 + |x-y|). \qquad \square \end{aligned}$$

Appendix C. Sobolev embedding. We summarize here some relevant results concerning embeddings of various Sobolev spaces (cf. [11]).

THEOREM C.1 (general Sobolev inequalities). *Let U be a bounded open subset of $\mathbb{R}^n$, with C^1 boundary. Assume $u \in W^{k,p}(U)$.*

(1) *If*

$$k < \frac{n}{p},$$

then $u \in L^q(U)$, where

$$\frac{1}{q} = \frac{1}{p} - \frac{k}{n}.$$

In addition,

$$\|u\|_{L^q(U)} \le C\|u\|_{W^{k,p}(U)}.$$

Here the constant C depends only on k, p, n, and U.

(2) *If*

$$k > \frac{n}{p},$$

then $u \in C^{k-[\frac{n}{p}]-1\gamma}(\bar{U})$, *where*

$$\gamma = \begin{cases} [\frac{n}{p}] + 1 - \frac{n}{p} & \text{if } \frac{n}{p} \text{ is not an integer,} \\ \text{any positive number} < 1 & \text{if } \frac{n}{p} \text{ is an integer.} \end{cases}$$

In addition,

$$\|u\|_{C^{k-[\frac{n}{p}]-1,\gamma}(\bar{U})} \leq C\|u\|_{W^{k,p}(U)},$$

where the constant C *depends only on* k, p, n, γ, *and* U.

REFERENCES

[1] A. Bensoussan and J.-L. Lions, *Impulse Control and Quasivariational Inequalities*, Heyden & Son, Philadelphia, 1984. Translation of *Contrôle Impulsionnel et Inéquations Quasi Variationnelles*, Gauthier-Villars, Paris, 1982.

[2] T. Bielecki and S. Pliska, *Risk sensitive asset management with transaction costs*, Finance Stoch., 4 (2000), pp. 1–33.

[3] A. L. Bronstein, L. P. Hughston, M. R. Pistorium, and M. Zervos, *Discretionary stopping of one-dimensional Itô diffusion with a staircase function*, J. Appl. Probab., 43 (2006), pp. 984–996.

[4] B. Bruder and H. Pham, *Impulse Control Problem on Finite Horizon with Execution Delay*, preprint, 2007. Available online at http://arxiv.org/abs/math/0703769.

[5] A. Cadenillas, T. Choulli, M. Taksar, and L. Zhang, *Classical and impulse stochastic control for the optimization of the dividend and risk policies of an insurance firm*, Math. Finance, 16 (2006), pp. 181–202.

[6] A. Cadenillas and F. Zapatero, *Optimal central bank intervention in the foreign exchange market*, J. Econom. Theory, 87 (1999), pp. 218–242.

[7] G. M. Constantinides and S. F. Richard, *Existence of optimal simple policies for discounted-cost inventory and cash management in continuous time*, Oper. Res., 26 (1978), pp. 620–636.

[8] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc., 27 (1992), pp. 1–62. Also available online at http://arxiv.org/PS_cache/math/pdf/9207/9207212v1.pdf.

[9] M. G. Crandall, R. T. Newcomb, and Y. Tomita, *Existence and uniqueness for viscosity solutions of degenerate quasilinear elliptic equations in* $\mathbb{R}^n$, Appl. Anal., 34 (1989), pp. 1–23.

[10] J. E. Eastham and K. J. Hastings, *Optimal impulse control of portfolios*, Math. Oper. Res., 13 (1988), pp. 588–605.

[11] L. C. Evans, *Partial Differential Equations*, AMS, Providence, RI, 1998.

[12] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, 2nd ed., Springer, New York, 2006.

[13] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer, New York, 1998.

[14] X. Guo and P. Tomecek, *Connections between singular control and optimal switching*, SIAM J. Control Optim., 47 (2008), pp. 421–443.

[15] J. M. Harrison, T. M. Sellke, and A. J. Taylor, *Impulse control of Brownian motion*, Math. Oper. Res., 8 (1983), pp. 454–466.

[16] J. M. Harrison and M. I. Taksar, *Instantaneous control of Brownian motion*, Math. Oper. Res., 8 (1983), pp. 439–453.

[17] H. Ishii, *On the equivalence of two notions of weak solutions, viscosity solutions and distribution solutions*, Funkcial. Ekvac., 38 (1995), pp. 101–120.

[18] M. J. Jeanblanc-Picqué and S. Shiryayev, *Optimization of the flow of dividends*, Russian Math. Surveys, 50 (1995), pp. 257–277.

[19] M. Jeanblanc-Picqué, *Impulse control method and exchange rate*, Math. Finance, 3 (1993), pp. 161–177.

[20] I. Karatzas and S. Shreve, *Methods of Mathematical Finance*, Springer-Verlag, New York, 1998.

[21] R. Korn, *Portfolio optimization with strictly positive transaction costs and impulse control*, Finance Stoch., 2 (1998), pp. 85–114.

[22] R. Korn, *Some applications of impulse control in mathematical finance*, Math. Meth. Oper. Res., 50 (1999), pp. 493–518.

[23] P. L. Lions, *Control of diffusion processes in R^N*, Comm. Pure Appl. Math., 34 (1981), pp. 121–147.

[24] P. L. Lions, *Viscosity solutions of fully nonlinear second-order equations and optimal stochastic control in infinite dimensions.* I. *The case of bounded stochastic evolutions*, Acta Math., 161 (1988), pp. 243–278.

[25] V. LyVath, M. Mnif, and H. Pham, *A model of optimal portfolio selection under liquidity risk and price impact*, Finance Stoch., 11 (2007), pp. 51–90.

[26] J. Ma, *On the principle of smooth fit for a class of singular stochastic control problems for diffusions*, SIAM J. Control Optim., 30 (1992), pp. 975–999.

[27] D. C. Mauer and A. Triantis, *Interactions of corporate financing and investment decisions: A dynamic framework*, J. Finance, 49 (1994), pp. 1253–1277.

[28] R. Merton, *Continuous-Time Finance*, Wiley-Blackwell, Oxford, UK, 1992.

[29] A. J. Morton and S. Pliska, *Optimal portfolio management with fixed transaction costs*, Math. Finance, 5 (1995), pp. 337–356.

[30] G. Mundaca and B. Øksendal, *Optimal stochastic intervention control with application to the exchange rate*, J. Math. Econom., 29 (1998), pp. 225–243.

[31] B. Øksendal and K. Reikvam, *Viscosity solutions of optimal stopping problems*, Stochastics Stochastics Rep., 62 (1998), pp. 285–301.

[32] B. Øksendal and A. Sulem, *Optimal consumption and portfolio with both fixed and proportional transaction costs*, SIAM J. Control Optim., 40 (2002), pp. 1765–1790.

[33] B. Øksendal and A. Sulem, *Optimal stochastic impulse control with delayed reaction*, Appl. Math. Optim., 58 (2008), pp. 243–255.

[34] M. Ormeci, J. G. Dai, and J. V. Vate, *Impulse control of Brownian motion: The constrained average cost case*, Oper. Res., 56 (2008), pp. 618–629.

[35] H. Pham, *On the smooth-fit property for one-dimensional optimal switching problem*, in Séminaire de Probabilités XL, Lecture Notes in Math. 1899, Springer, Berlin, 2007, pp. 187–199.

[36] M. Ramaswamy and S. Dharmatti, *Uniqueness of unbounded viscosity solutions for impulse control problem*, J. Math. Anal. Appl., 315 (2006), pp. 686–710.

[37] L. C. G. Rogers and D. Williams, *Diffusions, Markov Processes and Martingales*, Vol. II, Cambridge University Press, Cambridge, UK, 2000.

[38] A. Sulem, *A solvable one-dimensional model of a diffusion inventory system*, Math. Oper. Res., 11 (1986), pp. 125–133.

[39] M. I. Taksar, *Average optimal singular control and a related stopping problem*, Math. Oper. Res., 10 (1985), pp. 63–81.

[40] A. Triantis and J. E. Hodder, *Valuing flexibility as a complex option*, J. Finance, 45 (1990), pp. 549–565.

[41] A. Weerasinghe, *A bounded variation control problem for diffusion processes*, SIAM J. Control Optim., 44 (2006), pp. 389–417.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 2, pp. 618–640

© 2009 Society for Industrial and Applied Mathematics

OSCILLATORITY OF NONLINEAR SYSTEMS WITH STATIC FEEDBACK[*]

DENIS V. EFIMOV[†] AND ALEXANDER L. FRADKOV[‡]

Abstract. New Lyapunov-like conditions for oscillatority of dynamical systems in the sense of Yakubovich are proposed. Unlike previous results these conditions are applicable to nonlinear systems and allow for consideration of nonperiodic, e.g., chaotic modes. Upper and lower bounds for oscillations amplitude are obtained. The relation between the oscillatority bounds and excitability indices for the systems with the input are established. Control design procedure providing nonlinear systems with oscillatority property is proposed. Examples illustrating proposed results for Van der Pol system, Lorenz system, and Hindmarsh–Rose neuron model as well as computer simulation results are given.

Key words. analysis of oscillations, control of oscillations

AMS subject classifications. 34C15, 93B52

DOI. 10.1137/070706963

1. Introduction.

Most works on analysis or synthesis of nonlinear systems are devoted to studying stability-like behavior. Their typical results show that the motions of a system are close to a certain limit motion (limit mode) that either exists in the system or it is created by a controller. Evaluating deflection of the system trajectory from the limit mode, one may obtain quantitative information about system behavior [10, 27].

During recent years an interest in studying more complex dynamical systems behavior including oscillatory and, particularly, chaotic modes has grown significantly. Most authors deal with relaxed stability properties (orbital stability, Zhukovsky stability, partial stability) of some periodic limit modes [16, 19]. However, in order to study irregular, chaotic behavior the development of analysis and design methods for nonperiodical oscillations is needed. One such method based on the concept of excitability index (limit oscillation amplitude) for the systems excited with a bounded control was proposed in [7, 8].

It is worth noting that there exist many definitions for the term "oscillation" [11, 16]. For example, oscillation is understood as "any effect that varies in a back-and-forth or reciprocating manner" [6]. Otherwise, oscillation is the behavior of a sequence or a function, that does not converge, but also does not diverge to $+\infty$ or $-\infty$; that is, oscillation is the failure to have a limit [29]. Geometrically, an oscillating function of real numbers follows some path in a space, without settling into ever-smaller regions. In more simple cases the path might look like a loop coming back on itself, that is, periodic behavior; in more complex cases it may be a quite irregular movement covering a whole region [29]. Existing approaches based on Lyapunov stability theory [17, 23] or relaxed stability properties (orbital stability, Zhukovsky

[*] Received by the editors October 31, 2007; accepted for publication (in revised form) October 13, 2008; published electronically February 13, 2009.
http://www.siam.org/journals/sicon/48-2/70696.html

[†] Systems and Control, Université de Liège, Bat. 28, B-4000 Liege Sart-Tilman, Belgium (efimov@montefiore.ulg.ac.be) and Control of Complex Systems Laboratory, Institute of Problem of Mechanical Engineering, Bolshoi av., 61, V.O., St-Petersburg, 199178 Russia.

[‡] Control of Complex Systems Laboratory, Institute of Problem of Mechanical Engineering, Bolshoi av., 61, V.O., St-Petersburg, 199178 Russia (alf@control.ipme.ru).

stability, partial stability) [16, 19, 24] are not completely suitable for study of complex oscillations. Indeed, these approaches require information on some limit modes, which stability should be investigated (that is not suitable for chaotic or irregular oscillations, for example). Besides, these approaches are not suitable for distinguishing between simple bounded behavior and oscillating one (a trajectory can converge to a steady-state solution that is a stable behavior from any kind of stability definition, but it is not an oscillation). Despite significant success in study of regular oscillations [4, 5, 12, 18, 20], comprehensive solutions for generic irregular oscillations have not been obtained yet.

An important and useful concept for studying irregular oscillations is that of "oscillatority" introduced by V.A.Yakubovich in 1973 [31]. Frequency domain conditions for oscillatority were obtained for Lurie systems, and split in linear and nonlinear parts [16, 31, 32]. However, when studying physical and biological systems in many cases it is hard to decompose the system into two parts: Linear nominal system plus nonlinear feddback. Mechanical systems (where energy plays a role of Lyapunov function) serve as a widespread example of such systems. Extension of analysis and design methods to oscillations in such class of systems is still to appear.

In this paper an approach to detection of oscillations and design of oscillatory systems for a class of nonlinear systems is suggested. New conditions for oscillatority of dynamical systems in the sense of Yakubovich are proposed. These conditions are applicable to nonlinear systems, and they are formulated in terms of Lyapunov functions existence. As a result upper and lower bounds for oscillations amplitude are obtained. A variant of converse Lyapunov theorem for strictly unstable systems is proposed. The relation between the oscillatority bounds and excitability indices for the systems with input are established. Design procedure for oscillations excitation is presented. Potentiality of the proposed technique is illustrated by four examples of analytical computations and computer simulations.

The main advantage of the obtained solution consists in possibility of application to a wide range of oscillation analysis and design problems. The proposed conditions are applicable even in the cases when other existing solutions cannot be used due to complexity of oscillations or system models [5, 18, 20].

Section 2 contains auxiliary statements and definitions (two preliminary results are placed in Appendix). Main definitions and oscillation existence conditions are presented in section 3. Section 4 deals with the task of static feedback design, which ensures oscillations appearance in closed loop system with desired bounds on amplitude. Conclusion is given in section 5. Examples illustrating proposed results for Van der Pol system, Lorenz system, and Hindmarsh–Rose neuron model as well as computer simulation results are presented in the text.

2. Preliminaries.

Let us consider a general model of nonlinear dynamical system:

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}); \quad \mathbf{y} = \mathbf{h}(\mathbf{x}), \tag{1}$$

where $\mathbf{x} \in R^n$ is the state space vector; $\mathbf{u} \in R^m$ is the input vector; $\mathbf{y} \in R^p$ is the output vector; $\mathbf{f}$ and $\mathbf{h}$ are locally Lipschitz continuous functions on R^n, $\mathbf{h}(0) = 0$, and $\mathbf{f}(0, 0) = 0$. For initial condition $\mathbf{x}_0 \in R^n$ and Lebesgue measurable input $\mathbf{u}$ the solution $\mathbf{x}(\mathbf{x}_0, \mathbf{u}, t)$ of the system (1) is defined at least locally for $t \leq T$, $\mathbf{y}(\mathbf{x}_0, \mathbf{u}, t) = \mathbf{h}(\mathbf{x}(\mathbf{x}_0, \mathbf{u}, t))$ (further we will simply write $\mathbf{x}(t)$ or $\mathbf{y}(t)$ if all other arguments are clear from the context). If for all initial conditions $\mathbf{x}_0 \in R^n$ and inputs $\mathbf{u}$ the solutions are defined for all $t \geq 0$, then such system is called forward complete.

In this work we will consider feedback connection of system (1) with static system $\mathbf{u} = \mathbf{k}(\mathbf{y})$.

As usual, it is said that a continuous function $\rho : R_+ \to R_+$ belongs to class K, if it is strictly increasing and $\rho(0) = 0$; $\rho \in K_\infty$ if $\rho \in K$ and $\rho(s) \to \infty$ for $s \to \infty$; Lebesgue measurable function $\mathbf{x} : R_+ \to R^n$ is essentially bounded, if $\|\mathbf{x}\| = ess \sup \{ |\mathbf{x}(t)|, t \geq 0 \} < +\infty$, where $|\cdot|$ denotes usual Euclidean norm, $R_+ = \{ \tau \in R : \tau \geq 0 \}$. Notation $DV(\mathbf{x})\mathbf{F}(\cdot)$ stands for directional derivative of function V with respect to vector field $\mathbf{F}$ if function V is differentiable and for Dini derivative in the direction of $\mathbf{F}$

$$DV(\mathbf{x})\mathbf{F}(\cdot) = \lim_{t \to 0^+} \inf \frac{V(\mathbf{x} + t\mathbf{F}(\cdot)) - V(\mathbf{x})}{t}$$

if function V is Lipschitz continuous. In what follows we need the standard dissipativity property [30] and some its modifications. Function $f(x_1, \ldots, x_n)$ defined on R^n is called monotone if the condition $x_1 \leq x_1', \ldots, x_n \leq x_n'$ implies that everywhere either $f(x_1, \ldots, x_n) \leq f(x_1', \ldots, x_n')$ or $f(x_1, \ldots, x_n) \geq f(x_1', \ldots, x_n')$ everywhere.

Definition 1. *The system* (1) *is dissipative if there exists continuous function $V : R^n \to R_+$ and a function $\varpi : R^{n+m+p} \to R$ such that for all $\mathbf{x}_0 \in R^n$ and Lebesgue measurable and locally essentially bounded $\mathbf{u} : R_+ \to R^m$ the following inequality is satisfied:*

$$V(\mathbf{x}(t)) \leq V(\mathbf{x}_0) + \int_0^t \varpi(\mathbf{x}(\tau), \mathbf{y}(\tau), \mathbf{u}(\tau))\, d\tau, \quad t \geq 0. \tag{2}$$

The functions ϖ and V are called supply rate and storage functions of the system (1).

In the case when storage function is continuously differentiable, inequality (2) can be rewritten in a simple form:

$$\dot{V}(\mathbf{x}, \mathbf{u}) = L_{\mathbf{f}(\mathbf{x},\mathbf{u})} V(\mathbf{x}) \leq \varpi(\mathbf{x}, \mathbf{u}, \mathbf{y}).$$

Definition 2. *Dissipative system* (1) *is called*

– passive if $\varpi(\mathbf{x}, \mathbf{y}, \mathbf{u}) = \mathbf{y}^T\mathbf{u} - \beta(\mathbf{x})$, where β is a continuous function reflecting the dissipation rate in the system; if $\beta(\mathbf{x}) \geq \widehat{\beta}(|\mathbf{x}|), \widehat{\beta} \in K$, then system (1) *is called strictly passive* [13];

– h-dissipative, if it has continuously differentiable storage function V and

$$\underline{\alpha}(|\mathbf{y}|) \leq V(\mathbf{x}) \leq \overline{\alpha}(|\mathbf{x}|), \quad \omega(\mathbf{y}, \mathbf{u}) = -\alpha(|\mathbf{y}|) + \sigma(|\mathbf{u}|),$$
$$\sigma \in K, \quad \alpha, \underline{\alpha}, \overline{\alpha} \in K_\infty;$$

– input-output-to-state stable (IOSS), if it has continuously differentiable storage function W and [26]

$$\alpha_1(|\mathbf{x}|) \leq W(\mathbf{x}) \leq \alpha_2(|\mathbf{x}|), \quad \alpha_1, \alpha_2 \in K_\infty,$$
$$\omega(\mathbf{x}, \mathbf{y}, \mathbf{u}) = -\alpha_3(|\mathbf{x}|) + \sigma_1(|\mathbf{u}|) + \sigma_2(|\mathbf{y}|),$$

$\alpha_3 \in K_\infty, \quad \sigma_1, \sigma_2 \in K$ [26]*;*

– input-to-state stable (ISS), if it has continuously differentiable storage function U and [21]

$$\alpha_4(|\mathbf{x}|) \leq U(\mathbf{x}) \leq \alpha_5(|\mathbf{x}|), \quad \alpha_4, \alpha_5 \in K_\infty;$$

$$\omega(\mathbf{x}, \mathbf{y}, \mathbf{u}) = -\alpha_6(|\mathbf{x}|) + \delta(|\mathbf{u}|), \quad \alpha_6 \in K_\infty, \quad \delta \in K.$$

If inequality sign in (2) for the case $\varpi(\mathbf{x}, \mathbf{y}, \mathbf{u}) = \mathbf{y}^T\mathbf{u} - \beta(\mathbf{x})$ can be replaced with equality, then it is said that the system possesses passivity property with known dissipation rate β.

Term h-dissipativity was introduced with minor differences in [2]. An important example of such kind of systems is $\mathbf{y}$-strictly passive systems [13]. Also, passive system (1) can be transformed to h-dissipative under suitable feedback transformation.

Storage functions for IOSS and ISS systems are called Lyapunov functions [23, 26]. Existence of corresponding Lyapunov functions is the equivalent characterization of ISS and IOSS properties [21, 26].

The interrelations of the properties introduced in Definition 2 are established in the Lemma A.1 (see Appendix), which was proved in [1] with a more restrictive requirement for h-dissipativity storage function:

$$\alpha_7(|\mathbf{x}|) \leq V(\mathbf{x}) \leq \alpha_8(|\mathbf{x}|), \quad \alpha_7, \alpha_8 \in K_\infty.$$

General result in this direction was obtained in [15], where it was proven that input-to-output stability (this property is closely connected with h-dissipativity; see also [24] for more details) and IOSS are equivalent to ISS property for the system (1).

3. Oscillatority conditions. At first it is necessary to give a precise definition of the term "oscillatority" placed in the title of this section and the paper. There are several approaches to define oscillation phenomena for nonlinear dynamical systems [16]. Perhaps, the most general one is the concept introduced by Yakubovich [31, 32]. Here we recover definitions from [31, 32] with some mild modifications [11, 16] dealing with high dimension and general form of the system.

DEFINITION 3. *Solution* $\mathbf{x}(\mathbf{x}_0, 0, t)$ *with* $\mathbf{x}_0 \in R^n$ *of system* (1) *is called* $[\pi^-, \pi^+]$*-oscillation with respect to output* $\psi = \eta(\mathbf{x})$ *(where* $\eta : R^n \to R$ *is a continuous monotone function) if the solution is defined for all* $t \geq 0$ *and*

$$\underline{\lim}_{t \to +\infty} \psi(t) = \pi^-; \quad \overline{\lim}_{t \to +\infty} \psi(t) = \pi^+; \quad -\infty < \pi^- < \pi^+ < +\infty.$$

Solution $\mathbf{x}(\mathbf{x}_0, 0, t)$ *with* $\mathbf{x}_0 \in R^n$ *of system* (1) *is called* ***oscillating****, if there exist some output* ψ *and constants* π^-, π^+ *such that* $\mathbf{x}(\mathbf{x}_0, 0, t)$ *is* $[\pi^-, \pi^+]$*-oscillation with respect to the output* ψ*. Forward complete system* (1) *with* $\mathbf{u}(t) \equiv 0, t \geq 0$ *is called* ***oscillatory****, if for almost all* $\mathbf{x}_0 \in R^n$ *solutions of the system* $\mathbf{x}(\mathbf{x}_0, 0, t)$ *are oscillating. Oscillatory system* (1) *is called* ***uniformly oscillatory****, if for almost all* $\mathbf{x}_0 \in R^n$ *for corresponding solutions* $\mathbf{x}(\mathbf{x}_0, 0, t)$ *there exist output* ψ *and constants* π^-, π^+ *not depending on initial conditions.*

In other words, the solution $\mathbf{x}(\mathbf{x}_0, 0, t)$ is oscillating if output $\psi(t) = \eta(\mathbf{x}(\mathbf{x}_0, 0, t))$ is asymptotically bounded and there is no single limit value of $\psi(t)$ for $t \to +\infty$ that is close to definition of oscillatority from [29].

Note that the term "almost all solutions" is used to emphasize that generally system (1) for $\mathbf{u}(t) \equiv 0$, $t \geq 0$ has a nonempty set of equilibrium points; thus, there exists a set of initial conditions with zero measure such that corresponding solutions are not oscillations. It is worth stressing that constants π^- and π^+ are exact asymptotic bounds for output ψ. Therefore, in order to compute these values the exact estimates for the system solutions should be known, which is a hard task for general nonlinear system (1). Fortunately, information on approximate estimates of constants π^- and π^+ is sufficient to obtain estimates on system amplitude oscillations. The oscillation property introduced in Definition 3 is defined for zero input and any initial conditions of system (1). The following property is a closely related characterization of the system behavior, which develops the proposed above property for the case of nonzero input but for specified initial conditions [8].

Definition 4. *Let* $\mathbf{u} : R_+ \to R^m$ *be Lebesgue measurable and essentially bounded function and* $\mathbf{x}_0 \in R^n$ *be given such that* $\mathbf{x}(\mathbf{x}_0, \mathbf{u}, t)$ *be defined for all* $t \geq 0$. *The functions* $\chi^-_{\psi,\mathbf{x}_0}(\gamma), \chi^+_{\psi,\mathbf{x}_0}(\gamma)$ *defined for* $||\mathbf{u}|| \leq \gamma, \gamma \in R_+$ *are called lower and upper excitation indices of system* (1) *in point* $\mathbf{x}_0$ *with respect to the output* $\psi = \eta(\mathbf{x})$ *(where* $\eta : R^n \to R$ *is a continuous monotone function), if*

$$\left(\chi^-_{\psi,\mathbf{x}_0}(\gamma), \chi^+_{\psi,\mathbf{x}_0}(\gamma)\right) = \underset{(a,b)\in E(\gamma)}{\arg\max}\, \{b - a\},$$

$$E(\gamma) = \left\{ (a, b) : \begin{pmatrix} a = \underline{\lim}_{t\to+\infty}\, \eta(\mathbf{x}(\mathbf{x}_0, \mathbf{u}, t)), \\ b = \overline{\lim}_{t\to+\infty}\, \eta(\mathbf{x}(\mathbf{x}_0, \mathbf{u}, t)) \end{pmatrix} \right\}_{\|\mathbf{u}\|\leq\gamma}.$$

Lower and upper excitation indices of a forward complete system (1) *with respect to the output* ψ *are*

$$\chi^-_\psi(\gamma) = \inf_{\mathbf{x}_0\in R^n} \chi^-_{\psi,\mathbf{x}_0}(\gamma), \quad \chi^+_\psi(\gamma) = \sup_{\mathbf{x}_0\in R^n} \chi^+_{\psi,\mathbf{x}_0}(\gamma).$$

In the same way it is possible to introduce indices for a vector output $\psi = \eta(\mathbf{x})$, in this case indices would be vectors of the same dimension as the output ψ.

Excitation indices characterize ability of system (1) to exhibit forced or controllable oscillations caused by bounded inputs. It is clear that properties $\pi^- = \chi^-_\psi(0)$ and $\pi^+ = \chi^+_\psi(0)$ are satisfied. For nonzero inputs the excitability indices characterize maximum (over specified set of inputs $\|\mathbf{u}\| \leq \gamma$) asymptotic amplitudes $\chi^+_\psi(\gamma) - \chi^-_\psi(\gamma)$ of ψ.

Note that it is useful to calculate or estimate values of $\chi^-_\psi(\gamma)$ and $\chi^+_\psi(\gamma)$ for all $0 \leq \gamma < +\infty$ due to the following reason. Let oscillation amplitude be an inverse function of input amplitude, then the maximum oscillation amplitude be reached for some $\gamma*$ and for all $\gamma \geq \gamma*$ the amplitude decreases. The indices $\chi^-_\psi(\gamma)$ and $\chi^+_\psi(\gamma)$ preserve their values for $\gamma \geq \gamma*$. Hence, to catch the critical value $\gamma*$ of input amplitude providing maximum output amplitude for ψ, it is necessary to build full graphics of functions $\chi^-_\psi(\gamma)$ and $\chi^+_\psi(\gamma)$. The obtained characteristics will be closely related with the Cauchy gain recently investigated in [22] (in fact, $\pi^+ - \pi^-$ or $\chi^+_{\psi,\mathbf{x}_0}(\gamma) - \chi^-_{\psi,\mathbf{x}_0}(\gamma)$ are asymptotic amplitudes of $\psi(t)$ in the sense of [22] for zero or nonzero input $\mathbf{u}$, while $\chi^+_\psi(\gamma)$ reflects the Cauchy gain of the system (1)).

On the other hand, excitation indices from Definition 4 describe robustness of the oscillations property proposed in Definition 3. Conditions of oscillations existence in the system are summarized in the following theorem.

Theorem 1. *Let system* (1) *with* $\mathbf{u}(t) \equiv 0, t \in R_+, i.e.,$

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, 0), \tag{3}$$

have two continuous and locally Lipschitz Lyapunov functions V_1 *and* V_2 *satisfying for all* $\mathbf{x} \in R^n$ *the following inequalities:*

$$\upsilon_1(|\mathbf{x}|) \leq V_1(\mathbf{x}) \leq \upsilon_2(|\mathbf{x}|), \quad \upsilon_3(|\mathbf{x}|) \leq V_2(\mathbf{x}) \leq \upsilon_4(|\mathbf{x}|), \quad \upsilon_1, \upsilon_2, \upsilon_3, \upsilon_4 \in K_\infty,$$

and for some $0 < X_1 < \upsilon_1^{-1} \circ \upsilon_2 \circ \upsilon_3^{-1} \circ \upsilon_4(X_2) < +\infty$:

$DV_1(\mathbf{x})\,\mathbf{f}(\mathbf{x}, 0) > 0$ *for* $0 < |\mathbf{x}| < X_1$ *and* $\mathbf{x} \notin \Xi$,

$DV_2(\mathbf{x})\,\mathbf{f}(\mathbf{x}, 0) < 0$ *for* $|\mathbf{x}| > X_2$ *and* $\mathbf{x} \notin \Xi$, *where $\Xi \subset R^n$ is a set with zero Lebesgue measure, which contain all equilibriums of the system, and*

$$\Omega \cap \Xi = \emptyset,$$

where $\Omega = \{\, \mathbf{x} :\ \upsilon_2^{-1} \circ \upsilon_1(\, X_1\,) < |\, \mathbf{x}\,| < \upsilon_3^{-1} \circ \upsilon_4(\, X_2\,)\,\}$.

Then the system (3) *is oscillatory.*

Proof. Consider set $\Xi_0 \subset R^n$ of initial conditions not containing equilibrium points (which belong to set Ξ) of system (3). Then the solutions of the system starting from Ξ_0 are globally bounded, due to $\dot{V}_2 < 0$ for $|\, \mathbf{x}\,| > X_2$, and defined for all $t \geq 0$. Since the trajectory $\mathbf{x}(\, \mathbf{x}_0, 0, t\,)$, $\mathbf{x}_0 \in \Xi_0$, $t \geq 0$ is bounded, it has a nonempty closed, invariant, and compact ω-limit set, which belongs to the set Ω. Indeed, $V_2(\, t\,)$ asymptotically enters into the set where $V_2(\, t\,) < \upsilon_4(\, X_2\,)$, then $|\, \mathbf{x}(\, t\,)\,| < \upsilon_3^{-1} \circ \upsilon_4(\, X_2\,)$. In the same way function $V_1(\, t\,)$ is upper bounded and its limit values fall into the set where $V_1(\, t\,) > \upsilon_1(\, X_1\,)$; i.e., again $|\, \mathbf{x}(\, t\,)\,| > \upsilon_2^{-1} \circ \upsilon_1(\, X_1\,)$.

As it was supposed, Ω does not contain equilibrium points of the system. Hence, ω-limit set also does not include such invariant solutions. Then for each $\mathbf{x}_0 \in \Xi_0$ there exists an index i, $1 \leq i \leq n$ such that the solution is $[\, \pi^-, \pi^+\,]$-oscillation with respect to output x_i with $-\upsilon_3^{-1} \circ \upsilon_4(\, X_2\,) \leq \pi^- < \pi^+ < \upsilon_3^{-1} \circ \upsilon_4(\, X_2\,)$. Suppose that there is no such output. It means that for all $1 \leq i \leq n$ for output x_i equality $\pi^- = \pi^+$ holds. However, the latter could be true only in equilibrium points, which are excluded from the set Ω by the theorem conditions. Therefore, for almost all initial conditions the system solutions have such oscillating output and system (3) is oscillatory by Definition 3. Note that for different $\mathbf{x}_0 \in \Xi_0$ oscillating outputs x_i may exist for different i, $1 \leq i \leq n$. □

Remark 1. The set Ω determines lower and upper bounds for the values of π^- and π^+.

Like in [32] one can consider the Lyapunov function candidate for linearized near the origin system (3) as a function V_1 to prove local instability of the system. Instead of existence of storage function V_2, one can require just boundedness of the system solution $\mathbf{x}(\, t\,)$ with a known upper bound. It can be obtained using another approach not dealing with time derivative of Lyapunov function analysis. In this case Theorem 1 is transforming into Theorem 3.4 from [11]; see also [33].

COROLLARY 1. *Define Ξ as the set of the system* (3) *equilibriums, i.e.,* $\Xi = \{\, \mathbf{x} \in R^n : \mathbf{f}(\, \mathbf{x}, 0\,) = 0\,\}$, *which consists in isolated points, and* $\mathbf{A}(\, \mathbf{x}_0\,) = d\,\mathbf{f}(\, \mathbf{x}, 0\,)/d\,\mathbf{x}|_{\mathbf{x}=\mathbf{x}_0}$ *is the matrix of the system* (3) *linearization in point* $\mathbf{x}_0 \in R^n$. *Let the following conditions be valid:*

1. *For all $\mathbf{x}_0 \in \Xi$ the matrices of the system* (3) *linearization $\mathbf{A}(\, \mathbf{x}_0\,)$ have eigenvalues with positive real parts.*
2. *There exists $R > 0$ such that for almost all initial conditions $\mathbf{x}_0 \in R^n$:*

$$\lim_{t \to +\infty} |\mathbf{x}(\, \mathbf{x}_0, 0, t\,)| \leq R.$$

Then the system (3) *is oscillatory.*

Proof. By conditions of the corollary for almost all initial conditions the ω-limit set is compact and it does not contain the equilibriums of the system. Further the proof is similar to the proof of Theorem 1. □

Conditions of Theorem 1 are rather general and define the class of systems, which oscillatory behavior can be investigated by the approach, namely systems which have an attracting compact set in state space containing oscillatory movements of the systems. For such systems Theorem 1 or Corollary 1 give useful tools for testing their oscillating behavior and obtaining estimates for amplitude of oscillations.

Theorem 1 presents the sufficient conditions for system (1) to be oscillating in the sense of Yakubovich. It is possible to show that for a subclass of uniformly oscillating systems these conditions are also necessary. To prove this result we need the following two lemmas.

LEMMA 1. *Let there exist constant $r > 0$ such that for solutions of systems* (3) *the following property is satisfied:*

$$0 < |\mathbf{x}_0| < r \quad \Rightarrow \quad |\mathbf{x}(\mathbf{x}_0, 0, t)| > r$$

for all $t \geq T_{\mathbf{x}_0}$, where $0 < T_{\mathbf{x}_0} < +\infty$. Then there exists a continuous and locally Lipschitz–Lyapunov function $V_1(\mathbf{x})$ such that for all $\mathbf{x} \in R^n$

$$\upsilon_1(|\mathbf{x}|) \leq V_1(\mathbf{x}) \leq \upsilon_2(|\mathbf{x}|), \quad \upsilon_1, \upsilon_2 \in K_\infty,$$

additionally for all $0 < |\mathbf{x}| < r$ it holds:

$$D V_1(\mathbf{x}) \mathbf{f}(\mathbf{x}, 0) > 0.$$

Proof. For $|\mathbf{x}_0| < r$ let us introduce the function:

$$v(\mathbf{x}_0) = \inf_{0 \leq t \leq T_{\mathbf{x}_0}} |\mathbf{x}(\mathbf{x}_0, 0, t)|.$$

According to conditions of the lemma this function admits the following properties:

(i) $v(0) = 0$ and $v(\mathbf{x}) > 0$ for $0 < |\mathbf{x}| < r$;

(ii) $v(\mathbf{x}_0) = \inf_{0 \leq t \leq T_{\mathbf{x}_0} + \Delta} |\mathbf{x}(\mathbf{x}_0, 0, t)|$ for any $\Delta \geq 0$.

Additionally for $0 < |\mathbf{x}| < r$ the property $|v(0) - v(\mathbf{x})| = v(\mathbf{x}) \leq |\mathbf{x}| = |0 - \mathbf{x}|$ holds, which means continuity of function v at the origin. In the set $|\mathbf{x}| < r$ the relation $\delta(|\mathbf{x}|) \leq v(\mathbf{x}) \leq |\mathbf{x}|$ holds, where $\delta(s) = s(1+s)^{-1} \inf_{|\mathbf{x}| = s} v(\mathbf{x})$ is a continuous and strictly increasing function, $\delta(0) = 0$. The locally Lipschitz property of function v in the set $0 < |\mathbf{x}| < r$ follows from the following series of inequalities satisfied for any $\mathbf{x}_1$, $\mathbf{x}_2$ belonging to this set and some constants $L > 0$, $M > 0$, $T = \max\{T_{\mathbf{x}_1}, T_{\mathbf{x}_2}\}$:

$$|\mathbf{x}(\mathbf{x}_1, 0, t) - \mathbf{x}(\mathbf{x}_2, 0, t)| \leq M|\mathbf{x}_1 - \mathbf{x}_2|, \quad t \leq T;$$

$$||\mathbf{x}(\mathbf{x}_1, 0, t)| - |\mathbf{x}(\mathbf{x}_2, 0, t)|| \leq L|\mathbf{x}_1 - \mathbf{x}_2|, \quad t \leq T;$$

$$|v(\mathbf{x}_1) - v(\mathbf{x}_2)| = |\inf_{0 \leq t \leq T} |\mathbf{x}(\mathbf{x}_1, 0, t)| - \inf_{0 \leq t \leq T} |\mathbf{x}(\mathbf{x}_2, 0, t)||$$

$$\leq \sup_{0 \leq t \leq T} ||\mathbf{x}(\mathbf{x}_1, 0, t)| - |\mathbf{x}(\mathbf{x}_2, 0, t)|| \leq L|\mathbf{x}_1 - \mathbf{x}_2|.$$

By construction for initial conditions $|\mathbf{x}_0| < r$ the relation $v(\mathbf{x}(\mathbf{x}_0, 0, t)) \geq v(\mathbf{x}(\mathbf{x}_0, 0, 0))$, $t \leq T_{\mathbf{x}_0}$ holds, then $D v(\mathbf{x}) \mathbf{f}(\mathbf{x}, 0) \geq 0$ for all $|\mathbf{x}| < r$ and function $v(t)$ is not decreasing. To design a strictly increasing function let us introduce for $|\mathbf{x}_0| < r$ the function:

$$V_1(\mathbf{x}_0) = \inf_{0 \leq t \leq T_{\mathbf{x}_0}} k(t) v(\mathbf{x}(\mathbf{x}_0, 0, t)),$$

where $k : R_+ \to R_+$ is a continuously differentiable function with the following properties for all $t \in R_+$:

$$\kappa_1 \leq k(t) \leq \kappa_2, \quad 0 < \kappa_1 < \kappa_2 < +\infty; \quad \partial k / \partial t < 0.$$

As an example of such function k it is possible to choose the following one:

$$k(t) = \kappa_1 + (\kappa_2 - \kappa_1)\, e^{-t}, \quad \dot{k}(t) = (\kappa_1 - \kappa_2)\, e^{-t}.$$

By construction $V_1(0) = 0$ and $V_1(\mathbf{x}) > 0$ for $0 < |\mathbf{x}| < r$. In the set $|\mathbf{x}| < r$ the relation $\kappa_1\, \delta(|\mathbf{x}|) \le v(\mathbf{x}) \le \kappa_2 |\mathbf{x}|$ holds. The locally Lipschitz continuity of function V_1 in the set $0 < |\mathbf{x}| < r$ follows from the same arguments, since the following series of inequalities are satisfied for any $\mathbf{x}_1$, $\mathbf{x}_2$ belonging to this set and some constants $L > 0$, $M > 0$, $T = \max\{T_{\mathbf{x}_1}, T_{\mathbf{x}_2}\}$:

$$|\mathbf{x}(\mathbf{x}_1, 0, t) - \mathbf{x}(\mathbf{x}_2, 0, t)| \le M|\mathbf{x}_1 - \mathbf{x}_2|, \quad t \le T;$$

$$|v(\mathbf{x}_1) - v(\mathbf{x}_2)| \le L\,|\mathbf{x}_1 - \mathbf{x}_2|;$$

$$|v(\mathbf{x}(\mathbf{x}_1, 0, t)) - v(\mathbf{x}(\mathbf{x}_2, 0, t))| \le M\,L\,|\mathbf{x}_1 - \mathbf{x}_2|, \quad t \le T;$$

$$\begin{aligned} |V_1(\mathbf{x}_1) - V_1(\mathbf{x}_2)| = {} & |\inf_{0 \le t \le T_{\mathbf{x}_1}} k(t)\, v(\mathbf{x}(\mathbf{x}_1, 0, t)) - \inf_{0 \le t \le T_{\mathbf{x}_2}} k(t)\, v(\mathbf{x}(\mathbf{x}_2, 0, t))| \\ & \le \sup_{0 \le t \le T} k(t) |v(\mathbf{x}(\mathbf{x}_1, 0, t) - v(\mathbf{x}(\mathbf{x}_2, 0, t)| \le \kappa_2 M\,L\,|\mathbf{x}_1 - \mathbf{x}_2|. \end{aligned}$$

For $|\mathbf{x}| \ge r$ extend function $V_1 : R^n \to R_+$ in such a way that for all $\mathbf{x} \in R^n$ function V_1 is continuous and locally Lipschitz and there exist two functions $\upsilon_1, \upsilon_2 \in K_\infty$ such that for all $\mathbf{x} \in R^n$:

$$\upsilon_1(|\mathbf{x}|) \le V_1(\mathbf{x}) \le \upsilon_2(|\mathbf{x}|),$$

where $\upsilon_1(s) \le \kappa_1\, \delta(s)$, $\kappa_2\, s \le \upsilon_2(s)$ for $s < r$. By construction for initial conditions $0 < |\mathbf{x}_0| < r$ the following relations hold:

$$\begin{aligned} V_1(\mathbf{x}(\mathbf{x}_0, 0, t)) &= \inf_{0 \le \tau \le T_{\mathbf{x}(\mathbf{x}_0, 0, t)}} k(\tau)\, v(\mathbf{x}[\mathbf{x}(\mathbf{x}_0, 0, t), 0, \tau]) \\ &> \inf_{0 \le \tau \le T_{\mathbf{x}_0}} k(\tau)\, v(\mathbf{x}[\mathbf{x}_0, 0, \tau]) = V_1(\mathbf{x}_0), 0 < t \le T_{\mathbf{x}_0}, T_{\mathbf{x}(\mathbf{x}_0, 0, t)} < T_{\mathbf{x}_0}, \end{aligned}$$

then $D\,V_1(\mathbf{x})\,\mathbf{f}(\mathbf{x}, 0) > 0$ for all $0 < |\mathbf{x}| < r$. □

Under conditions of Lemma 1 solutions $\mathbf{x}(\mathbf{x}_0, 0, t)$ of the system (3) are locally unstable for initial conditions $\mathbf{x}_0$ which belong to the sphere $0 < |\mathbf{x}_0| < r$. According to the result of the lemma in this case the system (3) has corresponding Lyapunov function with positive time derivative for $0 < |\mathbf{x}| < r$. It is possible to say that Lemma 1 presents a variant of necessary conditions of a Lyapunov function existence for a subclass of strictly unstable systems, which is a new result.

LEMMA 2. *Let there exist constants $R > 0$ and $0 < T_{R,\mathbf{x}_0} < +\infty$ such that for solutions of the system* (3) *the following property is satisfied:*

$$|\mathbf{x}_0| > R \quad \Rightarrow \quad |\mathbf{x}(\mathbf{x}_0, 0, t)| < R, t \ge T_{R,\mathbf{x}_0}.$$

Then there exists a continuous and locally Lipschitz–Lyapunov function $V_2(\mathbf{x})$ such that for all $\mathbf{x} \in R^n$

$$\upsilon_3(|\mathbf{x}|) \le V_2(\mathbf{x}) \le \upsilon_4(|\mathbf{x}|), \quad \upsilon_3, \upsilon_4 \in K_\infty,$$

and for all $|\mathbf{x}| > R$ it holds that

$$D\,V_2(\mathbf{x})\,\mathbf{f}(\mathbf{x}, 0) < 0.$$

Proof. For $|\mathbf{x}_0| > R$ let us introduce the function

$$v(\mathbf{x}_0) = \sup_{t\geq 0} |\mathbf{x}(\mathbf{x}_0, 0, t)| = \sup_{T_{R,\mathbf{x}_0}\geq t\geq 0} |\mathbf{x}(\mathbf{x}_0, 0, t)|.$$

Under conditions of the lemma the property $v(\mathbf{x}) > R$ for $|\mathbf{x}| > R$ is satisfied. Additionally due to continuity of solutions of the system (3) with respect to initial conditions for each $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\begin{aligned}&\mathbf{x}_1 \in R^n, \quad \mathbf{x}_2 \in R^n,\\ &|\mathbf{x}_1 - \mathbf{x}_2| \leq \delta \quad \Rightarrow |\mathbf{x}(\mathbf{x}_2, 0, t) - \mathbf{x}(\mathbf{x}_1, 0, t)| \leq \varepsilon, t \leq t_{\max}, t_{\max} = \max\{T_{R,\mathbf{x}_1}, T_{R,\mathbf{x}_2}\}.\end{aligned}$$

Note that for solutions of the system the equality $\sup_{t_{\max}\geq t\geq 0} |\mathbf{x}(\mathbf{x}_i, 0, t)| = \sup_{t\geq 0} |\mathbf{x}(\mathbf{x}_i, 0, t)|$, $i = 1, 2$ is satisfied. Then for any initial conditions under constrain $|\mathbf{x}_1 - \mathbf{x}_2| \leq \delta$, $|\mathbf{x}_1| > R$, $|\mathbf{x}_2| > R$ it holds that

$$\begin{aligned}&|v(\mathbf{x}_1) - v(\mathbf{x}_2)|\\ &\quad = \left| \sup_{t_{\max}\geq t\geq 0} |\mathbf{x}(\mathbf{x}_1, 0, t)| - \sup_{t_{\max}\geq t\geq 0} |\mathbf{x}(\mathbf{x}_2, 0, t)| \right|\\ &\quad \leq \sup_{t_{\max}\geq t\geq 0} ||\mathbf{x}(\mathbf{x}_1, 0, t)| - |\mathbf{x}(\mathbf{x}_2, 0, t)|| \leq \varepsilon,\end{aligned}$$

which means continuity of function v for $|\mathbf{x}| > R$. In the set $|\mathbf{x}| > R$ for function v the following relation also holds:

$$|\mathbf{x}| \leq v(\mathbf{x}) \leq \delta(|\mathbf{x}|),$$

where $\delta(s) = s + \sup_{|\mathbf{x}|=s} v(\mathbf{x})$ is a continuous and strictly increasing function. The locally Lipschitz continuity of function v into set $|\mathbf{x}| > R$ follows from the series of inequalities satisfied for any $\mathbf{x}_1$, $\mathbf{x}_2$ from the set and some $L > 0$:

$$||\mathbf{x}(\mathbf{x}_1, 0, t)| - |\mathbf{x}(\mathbf{x}_2, 0, t)|| \leq L|\mathbf{x}_1 - \mathbf{x}_2|, \quad t \leq t_{\max},$$

$$\begin{aligned}&|v(\mathbf{x}_1) - v(\mathbf{x}_2)|\\ &\quad = \left| \sup_{t_{\max}\geq t\geq 0} |\mathbf{x}(\mathbf{x}_1, 0, t)| - \sup_{t_{\max}\geq t\geq 0} |\mathbf{x}(\mathbf{x}_2, 0, t)| \right|\\ &\quad \leq \sup_{t_{\max}\geq t\geq 0} ||\mathbf{x}(\mathbf{x}_1, 0, t)| - |\mathbf{x}(\mathbf{x}_2, 0, t)|| \leq L\,|\mathbf{x}_1 - \mathbf{x}_2|.\end{aligned}$$

By construction for all initial conditions with $|\mathbf{x}_0| > R$ it holds that

$$v(t) = v(\mathbf{x}(\mathbf{x}_0, 0, t)) \leq v(\mathbf{x}(\mathbf{x}_0, 0, 0)) = v(0),$$

then $D\,v(\mathbf{x})\,\mathbf{f}(\mathbf{x}, 0) \leq 0$ for $|\mathbf{x}| > R$ and function v is not increasing. To design a strictly decreasing function, consider the following one for $|\mathbf{x}_0| > R$:

$$V_2(\mathbf{x}_0) = \sup_{T_{R,\mathbf{x}_0}\geq t\geq 0} k(t)\,v(\mathbf{x}(\mathbf{x}_0, 0, t)),$$

where $k : R_+ \to R_+$ is a continuously differentiable function with properties for all $t \in R_+$:

$$\kappa_3 \leq k(t) \leq \kappa_4, \quad 0 < \kappa_3 < \kappa_4 < +\infty; \quad \partial k/\partial t > 0.$$

For example, it is possible to choose as a function $k(t)$ the following one:

$$k(t) = \frac{\kappa_3 + \kappa_4 t}{1+t}, \quad \dot{k}(t) = \frac{\kappa_4 - \kappa_3}{(1+t)^2}.$$

Under conditions of the lemma in the set $|\mathbf{x}| > R$ for function V_2 the relation $\kappa_3|\mathbf{x}| \le V_2(\mathbf{x}) \le \kappa_4\,\delta(|\mathbf{x}|)$ holds. For any initial conditions under constrain $|\mathbf{x}_1 - \mathbf{x}_2| \le \delta$, $|\mathbf{x}_1| > R$, $|\mathbf{x}_2| > R$ it holds that

$$\begin{aligned} &|V_2(\mathbf{x}_1) - V_2(\mathbf{x}_2)| \\ &\quad = \left| \sup_{T_{R,\mathbf{x}_1} \ge t \ge 0} k(t)\,v(\mathbf{x}(\mathbf{x}_1,0,t)) - \sup_{T_{R,\mathbf{x}_2} \ge t \ge 0} k(t)\,v(\mathbf{x}(\mathbf{x}_2,0,t)) \right| \\ &\quad \le \sup_{t_{\max} \ge t \ge 0} k(t)||\mathbf{x}(\mathbf{x}_1,0,t)| - |\mathbf{x}(\mathbf{x}_2,0,t)|| \le \kappa_4\,\varepsilon, \end{aligned}$$

which means continuity of function V_2 for $|\mathbf{x}| > R$. The locally Lipschitz continuity of function V_2 into set $|\mathbf{x}| > R$ follows from the same inequalities satisfied for any $\mathbf{x}_1$, $\mathbf{x}_2$ from the set and some $L > 0$:

$$\begin{aligned} &|V_2(\mathbf{x}_1) - V_2(\mathbf{x}_2)| \\ &\quad = \left| \sup_{T_{R,\mathbf{x}_1} \ge t \ge 0} k(t)\,v(\mathbf{x}(\mathbf{x}_1,0,t)) - \sup_{T_{R,\mathbf{x}_2} \ge t \ge 0} k(t)\,v(\mathbf{x}(\mathbf{x}_2,0,t)) \right| \\ &\quad \le \sup_{t_{\max} \ge t \ge 0} k(t)||\mathbf{x}(\mathbf{x}_1,0,t)| - |\mathbf{x}(\mathbf{x}_2,0,t)|| \le \kappa_4\,L|\,\mathbf{x}_1 - \mathbf{x}_2\,|. \end{aligned}$$

For $|\mathbf{x}| \le R$ we extend the definition of function V_2 such that for all $\mathbf{x} \in R^n$ function $V_2 : R^n \to R_+$ would be continuous and locally Lipschitz and for all $\mathbf{x} \in R^n$:

$$\upsilon_3(|\,\mathbf{x}\,|) \le V_2(\mathbf{x},t) \le \upsilon_4(|\,\mathbf{x}\,|),$$

where $\upsilon_3, \upsilon_4 \in K_\infty$ and $\kappa_4\,s \ge \upsilon_3(s)$, $\upsilon_4(s) \ge \kappa_3\,\delta(s)$ for $s > R$. By construction for all initial conditions with $|\mathbf{x}_0| > R$, it holds that

$$\begin{aligned} V_2(t) = V_2(\mathbf{x}(\mathbf{x}_0,0,t)) &= \sup_{T_{R,\mathbf{x}(\mathbf{x}_0,0,t)} \ge \tau \ge 0} k(\tau)\,v(\mathbf{x}[\mathbf{x}(\mathbf{x}_0,0,t),0,\tau]) \\ < \sup_{T_{R,\mathbf{x}_0} \ge \tau \ge 0} k(\tau)v(\mathbf{x}[\mathbf{x}_0,0,\tau]) &= V_2(\mathbf{x}_0) = V_2(0), 0 < t \le T_{R,\mathbf{x}_0}, T_{R,\mathbf{x}(\mathbf{x}_0,0,t)} < T_{R,\mathbf{x}_0}, \end{aligned}$$

and then $D\,V_2(\mathbf{x})\,\mathbf{f}(\mathbf{x},0) < 0$ for $|\mathbf{x}| > R$. □

Under conditions of the lemma set, $A = \{\mathbf{x} : |\mathbf{x}| < R\}$ is a globally attractive invariant set for solutions of system (3) with zero input; see also [17] for other converse Lyapunov theorems for set stability. Contrarily to the case considered in this paper, the Lyapunov functions $W : R^n \to R_+$ proposed in [17] possess for all $\mathbf{x} \in R^n$ the properties

$$\alpha_1(|\mathbf{x}|_A) \le W(\mathbf{x}) \le \alpha_2(|\mathbf{x}|_A), \quad \alpha_1, \alpha_2 \in K_\infty,$$

where $|\mathbf{x}|_A$ is the distance from point $\mathbf{x}$ to the set A, which stability is investigated.

Now we are ready to substantiate the necessary conditions of oscillatority.

THEOREM 2. *Let system* (3) *be uniformly oscillatory with respect to the output* $\psi = \eta(\mathbf{x})$ *(where* $\eta : R^n \to R$ *is a continuous function), and for all* $\mathbf{x} \in R^n$ *the following relations are satisfied:*

$$\chi_1(|\mathbf{x}|) \le \eta(\mathbf{x}) \le \chi_2(|\mathbf{x}|), \quad \chi_1, \chi_2 \in K_\infty;$$

the set of initial conditions for which the system is not oscillating consists in just one point $\Xi = \{\mathbf{x} : \mathbf{x} = 0\}$. *Then there exist two continuous and locally Lipschitz Lyapunov functions* $V_1 : R^n \to R_+$ *and* $V_2 : R^n \to R_+$ *such that for all* $\mathbf{x} \in R^n$ *the inequalities hold:*

$$\upsilon_1(|\mathbf{x}|) \le V_1(\mathbf{x}) \le \upsilon_2(|\mathbf{x}|), \quad \upsilon_3(|\mathbf{x}|) \le V_2(\mathbf{x}) \le \upsilon_4(|\mathbf{x}|), \quad \upsilon_1, \upsilon_2, \upsilon_3, \upsilon_4 \in K_\infty;$$

$$DV_1(\mathbf{x})\mathbf{f}(\mathbf{x}, 0) > 0 \text{ for } 0 < |\mathbf{x}| < \chi_2^{-1}(\pi^-);$$

$$DV_2(\mathbf{x})\mathbf{f}(\mathbf{x}, 0) < 0 \text{ for } |\mathbf{x}| > \chi_1^{-1}(\pi^+).$$

Proof. Since system (3) is uniformly oscillatory with respect to output $\psi = \eta(\mathbf{x})$, then for almost all initial conditions (except the origin) there exists constants $-\infty < \pi^- < \pi^+ < +\infty$ such that

$$\underline{\lim}_{t \to +\infty} \eta(\mathbf{x}(\mathbf{x}_0, 0, t)) = \underline{\lim}_{t \to +\infty} \psi(t) = \pi^-;$$
$$\overline{\lim}_{t \to +\infty} \eta(\mathbf{x}(\mathbf{x}_0, 0, t)) = \overline{\lim}_{t \to +\infty} \psi(t) = \pi^+.$$

By radial unboundedness and positive definiteness of function η it means that all solutions of the system converge to the invariant set $\Omega = \{\mathbf{x} : \chi_2^{-1}(\pi^-) \le \mathbf{x} \le \chi_1^{-1}(\pi^+)\}$. Then there exist constants $X_1 < \chi_2^{-1}(\pi^-)$ and $X_2 > \chi_1^{-1}(\pi^+)$ such that conditions of Lemmas 1 and 2 hold for $r = X_1$ and $R = X_2$. Based on these facts, the existence of Lyapunov functions V_1 and V_2 follows. □

For uniformly oscillatory systems with single equilibrium point at the origin, Theorems 1 and 2 give necessary and sufficient conditions of oscillations existence (Van der Pol or Hindmarsh and Rose systems (see below) are examples of uniformly oscillatory systems). The oscillatority concept introduced by Yakubovich covers situations of periodic and chaotic oscillations. That allows one to analyze behavior of wide spectrum of oscillating dynamical systems using common approach. Note that for chaotic systems constants π^- and π^+ evaluate geometrical size of strange attractor. Let us demonstrate on examples the efficiency of the proposed approach for analysis of oscillation phenomena in nonlinear systems.

Example 1. Consider the Van der Pol system:

$$\dot{x}_1 = x_2; \quad \dot{x}_2 = -x_1 + \varepsilon(1 - x_1^2)x_2,$$

where $\varepsilon > 0$ some parameter. To detect presence of oscillations in this system, it is required (according to Theorem 1) to find two Lyapunov functions, which establish local instability of equilibrium $(0, 0)$ and global boundedness of the system solutions. Since the system has only one equilibrium point in the origin, the set Ω from the theorem does not contain the point $(0, 0)$. Let us consider the following Lyapunov functions for $0 < \varepsilon \le 1$:

$$V_1(\mathbf{x}) = 0.5\left((1 - \varepsilon + \varepsilon^{-1})x_1^2 + (1 + \varepsilon^{-1})x_2^2 + \varepsilon(x_2 - \varepsilon x_1)^2\right);$$

$$V_2(\mathbf{x}) = 0.5\left(\varepsilon^{-1}x_2 - 2x_1 + 1/3\,x_1^3\right)^2 + 1/12\,x_1^4,$$

$$\dot{V}_1 = \varepsilon x_2^2 + (x_2 - \varepsilon x_1)^2 + \left[\varepsilon^3 x_1 - (1 + \varepsilon + \varepsilon^2)x_2\right] x_1^2 x_2;$$

$$\dot{V}_2 = -\left[0.5\sqrt{\varepsilon}\left(2 - \varepsilon^{-2}\right)x_1 - \varepsilon^{-0.5}x_2\right]^2 - 1/3\,\varepsilon^{-1}x_1^4$$
$$+ \left[0.25\,\varepsilon\left(2 - \varepsilon^{-2}\right)^2 + 2\,\varepsilon^{-1}\right] x_1^2.$$

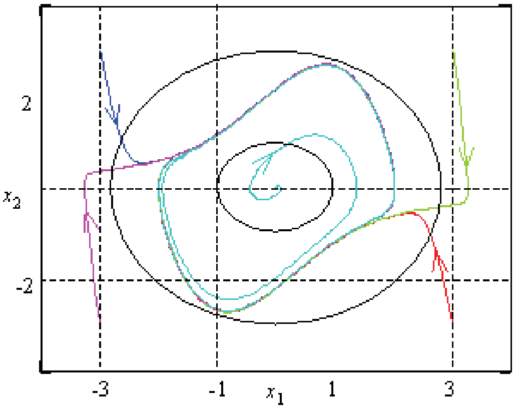


FIG. 1. *Trajectories and set* Ω *for Van der Pol system.*

Function $\dot{V}_1$ is strictly positive in the set $0 < |\mathbf{x}| < X_1$, where $X_1 = X_1(\varepsilon) > 0$ (the same conclusion was obtained in [12] for $\varepsilon = 1$, $X_1 = \sqrt{3}$). Instability of the system also can be verified for a linearized version of the system, which eigenvalues $\lambda_{1,2} = 0.5\left(\varepsilon \pm \sqrt{\varepsilon^2 - 4}\right)$ are always positive for $\varepsilon > 0$. Analyzing function $\dot{V}_2$ it is possible to obtain $X_2 \le \sqrt{3\,[\,0.25\,\varepsilon^2\,(\,2 - \varepsilon^{-2}\,)^2 + 2\,]}$. Results of the set Ω calculation and computer simulation of the system for $\varepsilon = 1$ are presented in Figure 1, where the set Ω is bounded by solid ellipses.

Example 2. Let us consider Lorenz model:

$$\begin{aligned}\dot{x} &= \sigma\,(\,y - x\,),\\ \dot{y} &= r\,x - y - x\,z,\\ \dot{z} &= -b\,z + x\,y\,,\end{aligned}$$

where parameters $\sigma = 10$, $r = 28$, and $b = 8/3$. With such choice of parameter values the system is chaotic, which is a good example of complex nonlinear oscillation processes. To apply the result of Theorem 1 here let us note that the system has three equilibriums with coordinates

$$\mathbf{x}_e^1 = (\,0\ 0\ 0\,)^T,\quad \mathbf{x}_e^2 = (\,\sqrt{72}\ \sqrt{72}\ 27\,)^T,\quad \mathbf{x}_e^3 = (\,-\sqrt{72}\ \ -\sqrt{72}\ 27\,)^T.$$

The matrix of linear approximation of this system at the equilibriums

$$A(\,\mathbf{x}_e\,) = \begin{bmatrix} -\sigma & \sigma & 0\\ r - \mathbf{x}_{e,3} & -1 & -\mathbf{x}_{e,1}\\ \mathbf{x}_{e,2} & \mathbf{x}_{e,1} & -b \end{bmatrix}$$

has for the given values of parameters eigenvalues with positive real parts for all equilibriums. Therefore the system is locally unstable. Lyapunov function

$$V(\,x,y,z\,) = 0.5\,\left(\,\sigma^{-1}x^2 + y^2 + (\,z - r\,)^2\,\right)$$

for this system has the following time derivative:

$$\begin{aligned}\dot{V} &= -x^2 + x\,y\, - y^2 - b\,z^2 + r\,b\,z\\ &\le -0.5\,x^2 - 0.5\,y^2 - 0.5\,b\,z^2 + 0.5\,b\,r^2\,,\end{aligned}$$

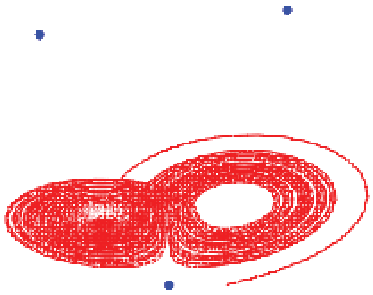


FIG. 2. *Trajectory of Lorenz system.*

which implies global boundedness of all trajectories of Lorenz system. All conditions of Corollary 1 are satisfied and system is oscillatory in the sense of Definition 3. An example of state space trajectory of the system is presented in Figure 2 (blue dots correspond to coordinates of equilibriums $\mathbf{x}_e^i$).

Example 3. A Hindmarsh and Rose model neuron is defined by the following system of differential equations [14]:

$$\begin{aligned}\dot{x} &= -a\,x^3 + b\,x^2 + y - z + u,\\ \dot{y} &= c - d\,x^2 - y,\\ \dot{z} &= \varepsilon\,[\,s\,(\,x - x_0\,) - z\,]\,,\end{aligned}$$

where $x \in R_+$ is the membrane potential, $y \in R_+$ is recovery variable, and $z \in R_+$ is adaptation variable. External stimulation is given by input $u \in R$. It is a well-known fact that this model demonstrates complex oscillatory behavior for the following values of the model parameters $a = 1$, $b = 3$, $c = 1$, $d = 5$, $s = 4$, $x_0 = 0.795$, $\varepsilon = 0.001$ with input $u = 0$. Let us investigate oscillatority property of the model for the case $u = 0$ applying the proposed approach.

As the first let us compute the number of equilibriums in the system which coordinates are solutions of the following system of nonlinear equations:

$$\begin{aligned}&-a\,x_e^3 + (\,b - d\,)\,x_e^2 - s\,x_e + s\,x_0 + c = 0\,;\\ &y_e = c - d\,x_e^2\,;\\ &z_e = s\,(\,x_e - x_0\,)\,.\end{aligned}$$

As in the first example we are interested in a situation when the model has a single equilibrium. This is the case when the first cubic equation above has only one real solution and two complex solutions. Under conditions

$$n \geq 0, \quad \frac{m}{6\,a} + \frac{2}{3}\,\frac{3\,s\,a - (\,b - d\,)^2}{a\,u} \neq 0,$$

$$n = 4\,s^3\,a - s^2\,(\,b - d\,)^2 + \left[\,27\,a^2\,(\,s\,x_0 + c\,) - 18\,s\,a\,(\,b - d\,) + 4\,(\,b - d\,)^3\,\right]\,(\,s\,x_0 + c\,),$$

$$m = \sqrt[3]{12\,a\,\sqrt{3\,n} - 36\,s\,a\,(\,b - d\,) + 108\,a^2\,(\,s\,x_0 + c\,) + 8\,(\,b - d\,)^3},$$

the model has the following single equilibrium

$$\begin{aligned}&x_e = a^{-1}\,(\,m/6 - 2/3\,[\,3\,s\,a - (\,b - d\,)^2\,]/m + (\,b - d\,)/3\,)\,;\\ &y_e = c - d\,x_e^2\,;\\ &z_e = s\,(\,x_e - x_0\,)\,.\end{aligned}$$

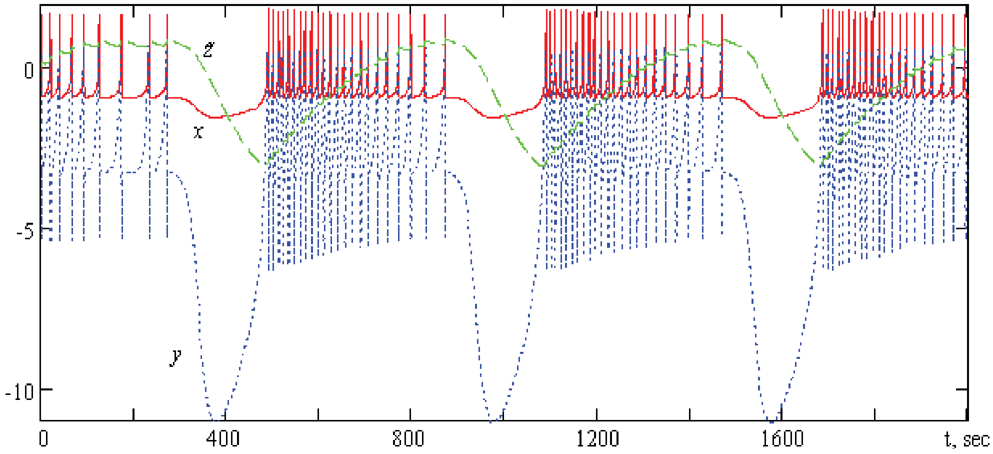


FIG. 3. *Trajectories of Hindmarsh and Rose neuron model.*

To prove global boundedness of the system solutions, it is possible to use the following Lyapunov function:

$$V_2 = 0.5\ \left(s\,x^2 + \varepsilon^{-1}z^2 + s\,a\,y^2/d^2\right),$$

in which the time derivative for the model admits inequality:

$$\dot{V}_2 \le s\,x\ \left(\ -0.5\,a\,x^3 + b\,x^2 + 8\,d^2x/a\ \right) - 0.25\,s\,a\,y^2/d^2 - 0.5\,z^2 + 8\,s\,a\,c^2/d^2 + 0.5\,s^2\,x_0^2.$$

To prove local instability of the equilibrium, consider linearization of the system with matrix

$$A(\,x_e, y_e, z_e\,) = \begin{bmatrix} -3\,a\,x_e^2 + 2\,b\,x_e & 1 & -1 \\ -2\,d\,x_e & -1 & 0 \\ \varepsilon\,s & 0 & -\varepsilon \end{bmatrix}.$$

According to Hurwitz criteria matrix A has eigenvalues with positive real parts if at least one from the following inequalities is satisfied:

$$3\,a\,x_e^2 - 2\,b\,x_e + 1 + \varepsilon \le 0, \quad 3\,a\,x_e^2 + 2\,(\,d - b\,)\,x_e + s \le 0,$$

$$3\,a\,(\,\varepsilon + 1\,)\,x_e^2 + 2\,(\,d - (\,\varepsilon + 1\,)\,b\,)\,x_e + \varepsilon\,(\,s + 1\,) \le 0,$$

$$9a^2(\varepsilon+1)x_e^4 + a[6d - 12(\varepsilon+1)b]x_e^3 + \left[4b[(\varepsilon+1)b - d] + 3a[\varepsilon^2 + (2+s)\varepsilon + 1]\right]x_e^2 \\ + 2\left[d - [\varepsilon^2 + (s+2)\varepsilon + 1]b\right]x_e + (s+1)\varepsilon^2 + \varepsilon \le 0.$$

Thus we obtain all set of restrictions on admissible values of the model parameters under which the system is uniformly oscillatory. The proposed values in [14] of the model parameters admit all these conditions (there exists single unstable equilibrium with globally bounded solutions). The result of the model simulation is shown in Figure 3, where $\tilde{z} = 10\,z$ is a scaled adaptation variable.

A link between oscillatority and excitation indices is established in the following corollary.

Corollary 2. *Let for initial condition* $\mathbf{x}_0 \in R^n$ *the solution* $\mathbf{x}(\mathbf{x}_0, \mathbf{k}(\mathbf{x}), t)$ *of system* (1) *with control* $\mathbf{u} = \mathbf{k}(\mathbf{x})$, $\mathbf{k}(0) = 0$ *be* $[\pi^-, \pi^+]$*-oscillation with respect to output*

$$\psi = \eta(\mathbf{x}), \quad \alpha_1(|\mathbf{x}|) \le \eta(\mathbf{x}), \quad \alpha_1 \in K_\infty.$$

Then excitation indices of system (1) *satisfy inequality*

$$\pi^+ - \pi^- \le \chi^+_{\psi,\mathbf{x}_0}(\gamma) - \chi^-_{\psi,\mathbf{x}_0}(\gamma),$$

for $\gamma \ge \gamma*$, *where* $\gamma* = \sup_{|\mathbf{x}| \le \alpha_1^{-1}(\pi^+)} |\mathbf{k}(\mathbf{x})|$.

Proof. From oscillatority property with respect to output ψ, the solutions of the closed by feedback $\mathbf{k}$ system (1) are asymptotically bounded:

$$|\mathbf{x}(t)| \le \alpha_1^{-1}(\pi^+), \quad t \ge 0.$$

Therefore input $\mathbf{u} = \mathbf{k}(\mathbf{x})$ is upper bounded by $\gamma \ge \gamma*$ and the statement follows from Definitions 3 and 4 (excitation indices are not decreasing functions of γ). □

Hence, to compute estimates on excitation indices it is enough to find some control $\mathbf{k}$ for system (1), which ensures oscillations existence in closed loop system.

In the proof of Theorem 1 a component of state space vector was proposed as an oscillating output. However, such output does not discover all features of oscillation processes in the system and it does not restrict the possible set of oscillating variables of the system. To avoid this obstacle we formulate the same conclusion for output oscillations of system (3) rewriting conditions of the theorem with respect to $\mathbf{y}$:

$$\upsilon_1(|\mathbf{y}|) \le V_1(\mathbf{x}) \le \upsilon_2(|\mathbf{y}|), \upsilon_3(|\mathbf{y}|) \le V_2(\mathbf{x}) \le \upsilon_4(|\mathbf{y}|),$$

$$DV_1(\mathbf{x})\mathbf{f}(\mathbf{x}, 0) > 0 \text{ for } 0 < |\mathbf{y}| < Y_1;$$

$$DV_1(\mathbf{x})\mathbf{f}(\mathbf{x}, 0) > 0 \text{ for } |\mathbf{y}| > Y_2,$$

$$Y_1 < \upsilon_1^{-1} \circ \upsilon_2 \circ \upsilon_3^{-1} \circ \upsilon_4(Y_2).$$

Then the set $\Omega = \{\mathbf{y} : \ \upsilon_2^{-1} \circ \upsilon_1(Y_1) < |\mathbf{y}| < \upsilon_3^{-1} \circ \upsilon_4(Y_2)\}$ and the system is oscillatory if set Ω does not contain equilibrium points of closed loop system $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, 0)$. A more constructive result, which points out on oscillating variables, can be presented as follows.

Lemma 3. *Let system* (1) *have IOSS Lyapunov function* W *and h-dissipative storage function* V *as in Definition* 2 *and* $\lim_{s\to+\infty} \alpha(s)^{-1}\sigma_2(s) < +\infty$ *(conditions of Lemma* A.1 *hold). Suppose that* $\mathbf{u} = \mathbf{k}(\mathbf{x})$ *and*

(i) $\alpha_6(|\mathbf{x}|) > \delta(|\mathbf{k}(\mathbf{x})|)$ *for* $|\mathbf{x}| > X \ge 0$ *and* $\mathbf{x} \notin \Xi$,

(ii) $L_{\mathbf{f}(\mathbf{x},\mathbf{k}(\mathbf{x}))}V(\mathbf{x}) > 0$ *for* $0 < |\mathbf{h}(\mathbf{x})| \le Y$ *and* $\mathbf{x} \notin \Xi$,

for some positive constants X *and* Y *with* $Y < \underline{\alpha}^{-1} \circ \overline{\alpha} \circ \alpha_4^{-1} \circ \alpha_5(X)$ *(where functions* $\alpha_4, \alpha_5, \alpha_6$ *and* δ *defined in Lemma* A.1*), set* Ξ *has zero Lebesgue measure. If set* $\Omega = \{V(\mathbf{x}) : \ \underline{\alpha}(Y) \le V(\mathbf{x}) \le \overline{\alpha} \circ \alpha_4^{-1} \circ \alpha_5(X)\}$ *does not contain equilibrium points of closed loop system* $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{k}(\mathbf{x}))$, *then the system is oscillatory.*

Proof. First of all note that from point (i) the system satisfies all conditions from Lemma A.1 to be ISS with respect to input $\mathbf{u}$ and it also has bounded (i.e., defined for all $t \ge 0$) solutions due to property (i). As before, $\mathbf{x}(t)$ and $\mathbf{y}(t)$ have nonempty closed and compact ω-limit sets, which are upper bounded by estimate $|\mathbf{x}| \le \alpha_4^{-1} \circ \alpha_5(X)$.

From point (ii) of the lemma it is possible to conclude that $\dot{V} > 0$ for small enough $0 < |\,\mathbf{y}\,| \le Y$. Then the set of ω-limit trajectories for function $V(\,t\,)$ belongs to the set Ω. Now the result immediately follows similarly to the final steps of Theorem 1 proof. □

Generically function V depends on part of variables only, which helps to define a subset of oscillating variables in the system. Additionally, Lemma 3 points out a way to find functions V_1 and V_2 ($V_1(\,\mathbf{x}\,) = V(\,\mathbf{x}\,)$ and $V_2(\,\mathbf{x}\,) = U(\,\mathbf{x}\,)$ from Appendix). Results of proposed theorems and Lemma 3 do not deal with feedback $\mathbf{k}$ design problem. Now let us continue with the task of control design that ensures desired oscillation parameters for passive systems.

4. Stabilization of oscillation regimes. In this section the problem of feedback design for passive system is considered, and the proposed feedback ensures oscillatority of closed loop system. Section 4 is based on result of Lemma A.2, although conditions imposed on feedback $\mathbf{k}$ in the Lemma A.2 look complex and hardly verified, they are very natural and can be easily resolved. For example, if σ_1 and σ_2 are quadratic functions of their arguments, then control $\mathbf{k}$ with linear growth rate with respect to $\mathbf{y}$ satisfies all proposed conditions.

THEOREM 3. *Let system* (1) *be passive with known dissipation rate β and IOSS in the sense of Definition* 2 *and*

$$\underline{\alpha}\,(\,|\,\mathbf{y}\,|) \le V(\,\mathbf{x}\,) \le \overline{\alpha}\,(\,|\,\mathbf{x}\,|), \quad \underline{\alpha}, \overline{\alpha} \in K_\infty.$$

Consider control $\mathbf{u} = \mathbf{k}(\,\mathbf{x}\,) + \mathbf{d}$, which possesses the following properties for all $\mathbf{x} \in R^n$:

(1) *for some $0 < K < +\infty$,*

$$|\,\mathbf{k}(\,\mathbf{x}\,)\,| \le \lambda\,(\,|\,\mathbf{y}\,|\,) + K;$$

(2) *decreasing of storage function V for large values of the output, i.e., inequality holds*

$$\beta(\,\mathbf{x}\,) - \mathbf{y}^T\mathbf{k}(\,\mathbf{x}\,) + \mu\,(\,|\,\mathbf{d}\,|\,) + \mu(\,K\,) \ge \kappa\,(\,|\,\mathbf{y}\,|\,) + \mathbf{y}^T\mathbf{d};$$

(3) $\mathbf{y}^T\mathbf{k}(\,\mathbf{x}\,) > \beta(\,\mathbf{x}\,)$ *for* $0 < |\,\mathbf{y}\,| < Y < +\infty$, $Y < \underline{\alpha}^{-1} \circ \overline{\alpha} \circ \alpha_4^{-1} \circ \alpha_5 \circ \alpha_6^{-1} \circ \delta(\,K\,)$, $\lim_{s \to +\infty} \frac{\sigma_2(\,s\,) + \sigma_1 \circ \lambda(\,s\,)}{\kappa(\,s\,)} < +\infty$, *where $\lambda \in K$, $\kappa \in K_\infty$, $\mu \in K$ (functions $\alpha_4, \alpha_5, \alpha_6$ and δ obtained in Lemma* A.2*) and $\mathbf{d} \in R^m$ is new input (Lebesgue measurable and essentially bounded function of time). Then*

(i) *system solutions are bounded;*

(ii) *if set $\Omega = \left\{ V(\,\mathbf{x}\,) :\ \underline{\alpha}(\,Y\,) \le V(\,\mathbf{x}\,) \le \overline{\alpha} \circ \alpha_4^{-1} \circ \alpha_5 \circ \alpha_6^{-1} \circ \delta(\,K\,) \right\}$ does not contain equilibrium points of system $\dot{\mathbf{x}} = \mathbf{f}\,(\,\mathbf{x}, \mathbf{k}(\,\mathbf{x}\,)\,)$ then for $\mathbf{d}(\,t\,) \equiv 0, t \ge 0$ closed loop system is an oscillatory one.*

Proof. Introduce partition of control input:

$$\mathbf{u} = \mathbf{k}(\,\mathbf{x}\,) = -\mathbf{k}_1(\,\mathbf{x}\,) + \mathbf{k}_2(\,\mathbf{x}\,),$$

such that

$$|\,\mathbf{k}_1(\,\mathbf{x}\,)\,| \le \lambda\,(\,|\,\mathbf{y}\,|\,), \quad |\,\mathbf{k}_2(\,\mathbf{x}\,)\,| \le K;$$

$$\mathbf{y}^T\mathbf{k}_1(\,\mathbf{x}\,) + \beta(\,\mathbf{x}\,) + \mu\,(\,|\,\mathbf{d}\,|\,) \ge \kappa\,(\,|\,\mathbf{y}\,|\,) + \mathbf{y}^T\mathbf{d};$$

$$\mathbf{y}^T\mathbf{k}_2(\,\mathbf{x}\,) > \beta(\,\mathbf{x}\,) + \mathbf{y}^T\mathbf{k}_1(\,\mathbf{x}\,) \text{ for } 0 < |\,\mathbf{y}\,| < Y < +\infty.$$

This separation is possible due to conditions of Theorem 3. Introduce auxiliary input $\tilde{\mathbf{d}} = \mathbf{d} + \mathbf{k}_2(\mathbf{x})$ (essentially bounded by conditions of the theorem $\|\tilde{\mathbf{d}}\| \le K + \|\mathbf{d}\|$). For system (1) all conditions of Lemma A.2 are satisfied for the feedback $\mathbf{u} = -\mathbf{k}_1(\mathbf{x}) + \tilde{\mathbf{d}}$ and system is ISS with respect to input $\tilde{\mathbf{d}}$. According to ISS property [21] and boundedness of $\tilde{\mathbf{d}}$, boundedness of system solution immediately follows and statement (i) of Theorem 3 is proven. To justify statement (ii) note that the conditions of Lemma 3 also hold. □

Theorem 3 extends the result from [3] and [28] to the case of general nonlinear dynamical systems. Additional special attention is given to the lower estimate of the oscillation amplitude for $\mathbf{d}(t) \equiv 0$, $t \ge 0$.

Exciting part $\mathbf{k}_2$ of feedback $\mathbf{k}$ defines the size of set Ω (due to constants Y and K are prescribed by $\mathbf{k}_2$) and, hence, it regulates the gap between values of π^- and π^+.

Remark 2. It is worth stressing that the control in Theorem 3 is proposed to satisfy some sector condition with respect to output $\mathbf{y}$. For design of such controls in practical application it is possible to use speed-gradient approach [9, 10], e.g., choose $\mathbf{u} = \varphi(\mathbf{y})$, where $\varphi(\mathbf{y})^T\mathbf{y} > 0$ for $0 < |\mathbf{y}| < Y_1$ and $\varphi(\mathbf{y})^T\mathbf{y} < 0$ for $|\mathbf{y}| > Y_2 > Y_1$.

Example 4. Let us consider controlled linear oscillator:

$$\dot{x}_1 = x_2; \quad \dot{x}_2 = -x_1 + u,$$

which is passive with storage function

$$V(\mathbf{x}) = 0.5\left(x_1^2 + x_2^2\right), \quad \dot{V} = x_2\, u,$$

and IOSS with corresponding Lyapunov function

$$W(\mathbf{x}) = 0.5\left(x_1^2 + (x_1 + x_2)^2\right),$$
$$\dot{W} \le -0.5\left(x_1^2 + x_2^2\right) + x_2^2 + u^2$$

with output $y = x_2$ ($\sigma_1(s) = \sigma_2(s) = s^2$). Then control $u = -k_1(\mathbf{x}) + k_2(\mathbf{x})$ with $k_1(\mathbf{x}) = a\,x_2$, $a > 0.5$ and $k_2(\mathbf{x}) = K\,sign(x_2)$ admits all condition of Theorem 3 with $\lambda(s) = a\,s$, $\kappa(s) = (a - 0.5)\,s^2$, $\mu(s) = 0.5\,s^2$. All functions σ_2, $\sigma_1 \circ \lambda$ and κ are square-law and, hence,

$$\lim_{s \to +\infty} \frac{\sigma_2(s) + \sigma_1 \circ \lambda(s)}{\kappa(s)} < +\infty;$$

inequality $x_2\, k_2(\mathbf{x}) > x_2\, k_1(\mathbf{x})$ holds for $0 < |x_2| < Y$, $Y = K/a$. This system is ISS for control $u = -k_1(\mathbf{x}) + d$ with ISS Lyapunov function:

$$U(\mathbf{x}) = W(\mathbf{x}) + \frac{1 + 2\,a^2}{a - 0.5}\,V(\mathbf{x}),$$
$$\dot{U} \le -0.5\left(x_1^2 + x_2^2\right) + \left(2 + \frac{0.5 + a^2}{a - 0.5}\right) d^2.$$

Then set

$$\Omega = \left\{\mathbf{x} :\; K/a \le |\mathbf{x}| \le \sqrt{1 + \frac{1.5\,a - 0.75}{a^2 + 0.5}}\,\sqrt{4 + \frac{1 + 2\,a^2}{a - 0.5}}\;K\right\}$$

is always nonempty. Simulation results and bounds of set Ω are shown in Figure 4 for $a = 1$ and $K = 1/3$.

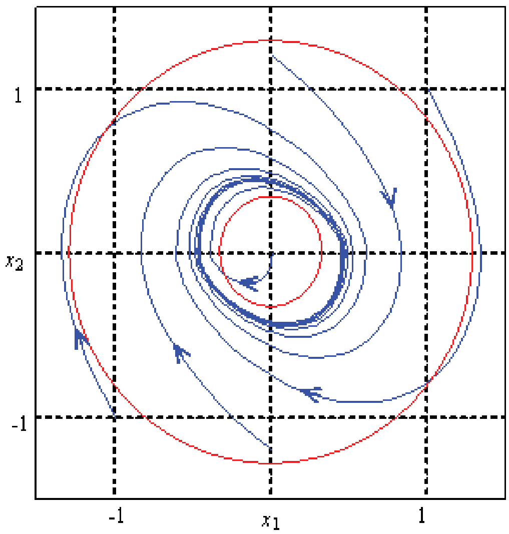


FIG. 4. *Trajectories of linear oscillator under nonlinear feedback.*

Based on the results of Theorem 3 and Corollary 2 it is possible to obtain the estimates of excitation indices of closed loop system for the case of nonvanishing signal $\mathbf{d}$.

COROLLARY 3. *Let all conditions of Theorem* 3 *hold. Then for* $\|\mathbf{d}\| \le \gamma < +\infty$

$$0 \le \chi_V^-(\gamma) \le \chi_V^+(\gamma) \le \overline{\alpha} \circ \alpha_4^{-1} \circ \alpha_5 \circ \alpha_6^{-1} \circ \delta(K+\gamma),$$

if additionally

$$\mathbf{y}(t)^T \mathbf{d}(t) \ge 0 \textit{ for all } t \ge 0, (3),$$

then

$$\underline{\alpha}(Y) \le \chi_V^-(\gamma) < \chi_V^+(\gamma) \le \overline{\alpha} \circ \alpha_4^{-1} \circ \alpha_5 \circ \alpha_6^{-1} \circ \delta(K+\gamma).$$

Proof. Upper estimate on excitation indices follows from ISS property of the system with respect to input $\tilde{\mathbf{d}}$ (asymptotic gain property in [25]). Now let us consider time derivative of storage function V:

$$\begin{aligned} \dot{V} &= \mathbf{y}^T(-\mathbf{k}_1(\mathbf{x}) + \mathbf{k}_2(\mathbf{x}) + \mathbf{d}) - \beta(\mathbf{x}) \\ &\ge \left[\mathbf{y}^T(-\mathbf{k}_1(\mathbf{x}) + \mathbf{k}_2(\mathbf{x})) - \beta(\mathbf{x})\right] + \mathbf{y}^T\mathbf{d}. \end{aligned}$$

From conditions of Theorem 3, the expression in square brackets is positive for $0 < |\mathbf{y}| < Y < +\infty$, but the presence of sign-varying term $\mathbf{y}^T\mathbf{d}$ allows one to claim only $0 \le \chi_V^-(\gamma) \le \chi_V^+(\gamma)$ in common case. But if $\mathbf{y}(t)^T\mathbf{d}(t) \ge 0$ for all $t \ge 0$, then

$$\begin{aligned} &[\mathbf{y}^T(-\mathbf{k}_1(\mathbf{x}) + \mathbf{k}_2(\mathbf{x})) - \beta(\mathbf{x})] + \mathbf{y}^T\mathbf{d} \\ &\quad \ge \mathbf{y}^T(-\mathbf{k}_1(\mathbf{x}) + \mathbf{k}_2(\mathbf{x})) - \beta(\mathbf{x}), \end{aligned}$$

and the desired result follows by the same line of consideration as in Theorem 3. Further let us suppose that it is possible a situation $\chi_V^-(\gamma) = \chi_V^+(\gamma)$ for some γ. But according to Definition 4, excitation indices admit conditions:

$$\gamma_1 \le \gamma_2 \quad \Rightarrow \quad \chi_V^-(\gamma_2) \le \chi_V^-(\gamma_1) \text{ and } \chi_V^+(\gamma_1) \le \chi_V^+(\gamma_2).$$

Applying the same arguments as in Corollary 2 for the results of Theorem 3 it is possible to obtain

$$0 < \chi_V^+(0) - \chi_V^-(0) \leq \overline{\alpha} \circ \alpha_4^{-1} \circ \alpha_5 \circ \alpha_6^{-1} \circ \delta(K) - \underline{\alpha}(Y),$$

therefore, $\chi_V^+(\gamma) - \chi_V^-(\gamma) > 0$ for any $\gamma \geq 0$. □

According to the corollary index $\chi_V^+(\gamma)$ is always bounded, that is more, it can not be equal to $\chi_V^-(\gamma)$ for any $\gamma \in R_+$ with (3). Thus, system can not lose its oscillation ability for any large enough input disturbance possessing "coordination" condition (3) and such input $\mathbf{d}$ does not provide new equilibrium points into set $\Omega = \left\{ V(\mathbf{x}) : \underline{\alpha}(Y) \leq V(\mathbf{x}) \leq \overline{\alpha} \circ \alpha_4^{-1} \circ \alpha_5(K + \gamma) \right\}$ for system $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{k}_1(\mathbf{x}) + \mathbf{k}_2(\mathbf{x}) + \mathbf{d})$. Also it is worth to note, that the requirement (3) can be satisfied for $t \geq T$ only, where $0 \leq T < +\infty$.

5. Conclusion. In this paper conditions for oscillatority in the sense of Yakubovich applicable to nonlinear systems are proposed. Upper and lower bounds for oscillation amplitude are evaluated. Presented conditions are also necessary for some special class of uniformly oscillating systems. Relation between the oscillatority bounds and excitability indices for the systems with input is established. An important advantage of the results of the paper is their applicability to complex nonperiodic (e.g., chaotic) oscillations. Such an advantage is achieved due to using the concept of oscillatority in the sense of Yakubovich as the starting point of the whole study. The results are illustrated by examples: Evaluation of oscillations for Van der Pol and Hindmarsh–Rose neuron systems. As a side result a smooth nonquadratic Lyapunov function providing boundedness of Van der Pol system solutions has been found.

Appendix.

LEMMA A.1. *Let system* (1) *have IOSS Lyapunov function W and h-dissipative storage function V as in Definition* 2. *If*

$$\lim_{s \to +\infty} \frac{\sigma_2(s)}{\alpha(s)} < +\infty,$$

then system (1) *is ISS with ISS Lyapunov function*

$$U(\mathbf{x}) = V(\mathbf{x}) + \tilde{W}(\mathbf{x}), \quad \tilde{W}(\mathbf{x}) = \rho(W(\mathbf{x})),$$
$$\rho(r) = \int_0^r q(s)\, ds, \quad q(s) = \frac{\alpha \circ \sigma_2^{-1}\left(0.25\,\alpha_3 \circ \alpha_2^{-1}(s)\right)}{1 + 0.5\,\alpha_3 \circ \alpha_2^{-1}(s)},$$
$$\alpha_4(s) = \rho \circ \alpha_1(s), \quad \alpha_5(s) = \overline{\alpha}(s) + \rho \circ \alpha_2(s), \quad \alpha_6(s) = 0.5\, q(\alpha_1(s))\, \alpha_3(s),$$
$$L_{\mathbf{f}(\mathbf{x},\mathbf{u})} U(\mathbf{x}) \leq -\alpha_6(|\mathbf{x}|) + \delta(|\mathbf{u}|), \quad \delta(s) = \sigma(s) + 2\,\chi(2\,\sigma_1(s))\,\sigma_1(s),$$
$$\chi(2\,\sigma_2(s)) = \alpha(s)\,[1 + 2\,\sigma_2(s)]^{-1}.$$

Proof. According to conditions of the lemma and Definition 2, the following series of inequalities holds for all $\mathbf{x} \in R^n$ and $\mathbf{u} \in R^m$:

$$\alpha_1(|\mathbf{x}|) \leq W(\mathbf{x}) \leq \alpha_2(|\mathbf{x}|); \quad L_{\mathbf{f}(\mathbf{x},\mathbf{u})} W(\mathbf{x}) \leq -\alpha_3(|\mathbf{x}|) + \sigma_1(|\mathbf{u}|) + \sigma_2(|\mathbf{y}|);$$
$$\underline{\alpha}(|\mathbf{y}|) \leq V(\mathbf{x}) \leq \overline{\alpha}(|\mathbf{x}|); \quad L_{\mathbf{f}(\mathbf{x},\mathbf{u})} V(\mathbf{x}) \leq -\alpha(|\mathbf{y}|) + \sigma(|\mathbf{u}|),$$

where $\alpha, \alpha_1, \alpha_2, \alpha_3, \underline{\alpha}, \overline{\alpha} \in K_\infty$ and $\sigma, \sigma_1, \sigma_2 \in K$. Let us consider a new IOSS Lyapunov function

$$\tilde{W}(\mathbf{x}) = \rho(W(\mathbf{x})), \quad \rho(r) = \int_0^r q(s)\, ds,$$

where q is some function from class K (that will be defined later). Clearly function $\tilde{W}$ is again continuously differentiable, positive definite, and radially unbounded provided that $\rho \in K_\infty$. Its time derivative admits an estimate:

$$L_{\mathbf{f}(\mathbf{x},\mathbf{u})}\tilde{W}(\mathbf{x}) \le q(W(\mathbf{x}))\,[-\alpha_3(|\mathbf{x}|) + \sigma_1(|\mathbf{u}|) + \sigma_2(|\mathbf{y}|)]\,.$$

To disclose the above inequality let us analyze consequently three situations:

(a) If $0.5\,\alpha_3(|\mathbf{x}|) \ge \sigma_1(|\mathbf{u}|) + \sigma_2(|\mathbf{y}|)$, then

$$L_{\mathbf{f}(\mathbf{x},\mathbf{u})}\tilde{W}(\mathbf{x}) \le -0.5\,q(W(\mathbf{x}))\,\alpha_3(|\mathbf{x}|)\,;$$

(b) If $0.5\,\alpha_3(|\mathbf{x}|) < \sigma_1(|\mathbf{u}|) + \sigma_2(|\mathbf{y}|)$ and $\sigma_1(|\mathbf{u}|) \le \sigma_2(|\mathbf{y}|)$, then

$$\begin{aligned} L_{\mathbf{f}(\mathbf{x},\mathbf{u})}\tilde{W}(\mathbf{x}) &\le -q(W(\mathbf{x}))\,\alpha_3(|\mathbf{x}|) + 2\,q(W(\mathbf{x}))\,\sigma_2(|\mathbf{y}|) \\ &\le -q(W(\mathbf{x}))\,\alpha_3(|\mathbf{x}|) + 2\,\chi(2\,\sigma_2(|\mathbf{y}|))\,\sigma_2(|\mathbf{y}|)\,, \end{aligned}$$

where $\chi(s) = q \circ \alpha_2 \circ \alpha_3^{-1}(2\,s)$;

(c) If $0.5\,\alpha_3(|\mathbf{x}|) < \sigma_1(|\mathbf{u}|) + \sigma_2(|\mathbf{y}|)$ and $\sigma_1(|\mathbf{u}|) > \sigma_2(|\mathbf{y}|)$, then

$$\begin{aligned} L_{\mathbf{f}(\mathbf{x},\mathbf{u})}\tilde{W}(\mathbf{x}) &\le -q(W(\mathbf{x}))\,\alpha_3(|\mathbf{x}|) + 2\,q(W(\mathbf{x}))\,\sigma_1(|\mathbf{u}|) \\ &\le -q(W(\mathbf{x}))\,\alpha_3(|\mathbf{x}|) + 2\,\chi(2\,\sigma_1(|\mathbf{u}|))\,\sigma_1(|\mathbf{u}|)\,. \end{aligned}$$

Thus, the time derivative of function $\tilde{W}$ calculated for system (1) can be rewritten in the form:

$$\begin{aligned} L_{\mathbf{f}(\mathbf{x},\mathbf{u})}\tilde{W}(\mathbf{x}) &\le -0.5\,q(W(\mathbf{x}))\,\alpha_3(|\mathbf{x}|) \\ &\quad + 2\,\chi(2\,\sigma_2(|\mathbf{y}|))\,\sigma_2(|\mathbf{y}|) + 2\,\chi(2\,\sigma_1(|\mathbf{u}|))\,\sigma_1(|\mathbf{u}|)\,. \end{aligned}$$

Let function χ be taken to possess the following equality:

$$\chi(2\,\sigma_2(s)) = \frac{\alpha(s)}{1 + 2\,\sigma_2(s)},$$

such choice of χ is possible due to

$$\lim_{s \to +\infty} \frac{\sigma_2(s)}{\alpha(s)} < +\infty$$

with $q(s) = \frac{\alpha \circ \sigma_2^{-1}\left(0.25\,\alpha_3 \circ \alpha_2^{-1}(s)\right)}{1 + 0.5\,\alpha_3 \circ \alpha_2^{-1}(s)}$ from class K. Then system (1) is ISS with ISS Lyapunov function $U(\mathbf{x}) = V(\mathbf{x}) + \tilde{W}(\mathbf{x})$ ($\alpha_4(s) = \rho \circ \alpha_1(s)$, $\alpha_5(s) = \overline{\alpha}(s) + \rho \circ \alpha_2(s)$), indeed:

$$\begin{aligned} L_{\mathbf{f}(\mathbf{x},\mathbf{u})}U(\mathbf{x}) &\le -0.5\,q(W(\mathbf{x}))\,\alpha_3(|\mathbf{x}|) + \sigma(|\mathbf{u}|) \\ &\quad + 2\,\chi(2\,\sigma_1(|\mathbf{u}|))\,\sigma_1(|\mathbf{u}|) \le -\alpha_6(|\mathbf{x}|) + \delta(|\mathbf{u}|)\,, \end{aligned}$$

where $\alpha_6(s) = 0.5\,q(\alpha_1(s))\,\alpha_3(s)$ and $\delta(s) = \sigma(s) + 2\,\chi(2\,\sigma_1(s))\,\sigma_1(s)$. □

The next lemma is a corollary of Lemma A.1 presenting a variant of ISS stabilizing control law for a passive system.

LEMMA A.2. *Let system* (1) *be passive and IOSS in the sense of Definition* 2 *and*

$$\underline{\alpha}(|\mathbf{y}|) \le V(\mathbf{x}) \le \overline{\alpha}(|\mathbf{x}|), \quad \underline{\alpha}, \overline{\alpha} \in K_\infty.$$

Then control

$$\mathbf{u} = -\mathbf{k}(\mathbf{x}) + \mathbf{d}, \quad |\mathbf{k}(\mathbf{x})| \le \lambda(|\mathbf{y}|), \quad \lambda \in K;$$
$$\mathbf{y}^T\mathbf{k}(\mathbf{x}) + \beta(\mathbf{x}) \ge \kappa(|\mathbf{y}|) + 0.5\,|\mathbf{y}|^2, \quad \kappa \in K_\infty;$$
$$\lim_{s\to+\infty} \frac{\sigma_2(s) + \sigma_1 \circ \lambda(s)}{\kappa(s)} < +\infty,$$

where $\mathbf{d} \in R^m$ *is new input (Lebesgue measurable and essentially bounded function of time), and provides for the system ISS property with ISS Lyapunov function:*

$$U(\mathbf{x}) = V(\mathbf{x}) + \tilde{W}(\mathbf{x}), \quad \tilde{W}(\mathbf{x}) = \rho(W(\mathbf{x})), \quad \rho(r) = \int_0^r q(s)\,ds,$$
$$q(s) = \frac{\kappa \circ \tilde{\sigma}_2^{-1}\left(0.25\,\alpha_3 \circ \alpha_2^{-1}(s)\right)}{1 + 0.5\,\alpha_3 \circ \alpha_2^{-1}(s)}, \quad \alpha_4(s) = \rho \circ \alpha_1(s),$$
$$\alpha_5(s) = \overline{\alpha}(s) + \rho \circ \alpha_2(s), \quad \alpha_6(s) = 0.5\,q(\alpha_1(s))\,\alpha_3(s),$$
$$\delta(s) = 0.5\,s^2 + 2\,\chi(2\,\sigma_1(2\,s))\,\sigma_1(2\,s).$$

Proof. From Definition 2 the following conditions hold for all $\mathbf{x} \in R^n$ and $\mathbf{u} \in R^m$:

$$\alpha_1(|\mathbf{x}|) \le W(\mathbf{x}) \le \alpha_2(|\mathbf{x}|);$$
$$L_{\mathbf{f}(\mathbf{x},\mathbf{u})}W(\mathbf{x}) \le -\alpha_3(|\mathbf{x}|) + \sigma_1(|\mathbf{u}|) + \sigma_2(|\mathbf{y}|);$$
$$\underline{\alpha}(|\mathbf{y}|) \le V(\mathbf{x}) \le \overline{\alpha}(|\mathbf{x}|); \quad L_{\mathbf{f}(\mathbf{x},\mathbf{u})}V(\mathbf{x}) \le -\beta(|\mathbf{x}|) + \mathbf{y}^T\mathbf{u}$$

with $\alpha_1, \alpha_2, \alpha_3, \underline{\alpha}, \overline{\alpha} \in K_\infty$, $\sigma_1, \sigma_2 \in K$ and β some nonnegative definite function. Substituting control in these inequalities, it is possible to obtain

$$\begin{aligned} L_{\mathbf{f}(\mathbf{x},\mathbf{u})}W(\mathbf{x}) &\le -\alpha_3(|\mathbf{x}|) + \sigma_1(|\mathbf{d} - \mathbf{k}(\mathbf{x})|) + \sigma_2(|\mathbf{y}|) \\ &\le -\alpha_3(|\mathbf{x}|) + \sigma_1(2\,|\mathbf{d}|) + \sigma_1(2\,\lambda(|\mathbf{y}|)) + \sigma_2(|\mathbf{y}|); \\ L_{\mathbf{f}(\mathbf{x},\mathbf{u})}V(\mathbf{x}) &\le -\beta(|\mathbf{x}|) + \mathbf{y}^T(\mathbf{d} - \mathbf{k}(\mathbf{x})) \le -\kappa(|\mathbf{y}|) + 0.5\,|\mathbf{d}|^2. \end{aligned}$$

Thus, such control provides for closed loop system IOSS property and h-dissipativity with respect to new input $\mathbf{d}$. If

$$\lim_{s\to+\infty} \frac{\tilde{\sigma}_2(s)}{\kappa(s)} < +\infty, \quad \tilde{\sigma}_2(s) = \sigma_2(s) + \sigma_1 \circ \lambda(s),$$

then all conditions of Lemma A1 are satisfied and the system is ISS with ISS Lyapunov function

$$U(\mathbf{x}) = V(\mathbf{x}) + \tilde{W}(\mathbf{x}), \quad \tilde{W}(\mathbf{x}) = \rho(W(\mathbf{x})), \quad \rho(r) = \int_0^r q(s)\,ds,$$
$$q(s) = \frac{\kappa \circ \tilde{\sigma}_2^{-1}\left(0.25\,\alpha_3 \circ \alpha_2^{-1}(s)\right)}{1 + 0.5\,\alpha_3 \circ \alpha_2^{-1}(s)}, \quad \alpha_4(s) = \rho \circ \alpha_1(s),$$
$$\alpha_5(s) = \overline{\alpha}(s) + \rho \circ \alpha_2(s), \quad \alpha_6(s) = 0.5\,q(\alpha_1(s))\,\alpha_3(s),$$
$$\delta(s) = 0.5\,s^2 + 2\,\chi(2\,\sigma_1(2\,s))\,\sigma_1(2\,s). \qquad \square$$

REFERENCES

[1] D. Angeli, *Input-to-state stability of PD-controlled robotic systems*, Automatica, 35 (1999), pp. 1285–1290.

[2] D. Angeli, E.D. Sontag, and Y.A. Wang, *A characterization of integral input to state stability*, Systems Control Lett., 38 (1999), pp. 209–217.

[3] M. Arcak and A. Teel, *Input-to-state stability for a class of Lurie systems*, Automatica, 38 (2002), pp. 1945–1949.

[4] V.I. Arnold, V.V. Kozlov, and A.I. Neishtadt, *Mathematical Aspects of Classical and Celestial Mechanics*, Springer Verlag, Berlin, 1997.

[5] N.N. Bogolyubov and Y.A. Mitropolsky, *Asymptotic Methods in the Theory of Nonlinear Oscillations*, Gordon and Breach, New York, 1962.

[6] *Encyclopedia of Science and Technology*, McGraw-Hill, 2005.

[7] A.L. Fradkov, *Exploring nonlinearity by feedback*, Phys. D., 128 (1999), pp. 159–168.

[8] A.L. Fradkov, *Investigation of physical systems by means of feedback*, Automat. Remote Control, 60 (1999), pp. 3–22.

[9] V.N. Fomin, A.L. Fradkov, and V.A. Yakubovich, *Adaptive Control of Dynamical Plants*, Nauka, Moscow, 1981 (in Russian).

[10] A.L. Fradkov, I.V. Miroshnik, and V.O. Nikiforov, *Nonlinear and Adaptive Control of Complex Systems*, Kluwer Academic Publishers, 1999.

[11] A.L. Fradkov and A.Yu. Pogromsky, *Introduction to Oscillations and Chaos*, World Scientific, Singapore, 1998.

[12] C. Hayachi, *Nonlinear Oscillations in Physical Systems*, McGraw-Hill Book Company, New York, 1964.

[13] D. Hill and P. Moylan, *Dissipative dynamical systems: Basic input – output and state properties*, J. Franklin Inst., 309 (1980), pp. 327–357.

[14] J.L. Hindmarsh and R.M. Rose, *A model of neunoral bursting using* 3 *coupled* 1*st order differential-equations*, Proc. R. Soc. Lond., B 221 (1984), pp. 87–102.

[15] Z.-P. Jiang, A. Teel, and L. Praly, *Small – gain theorem for ISS systems and applications*, Math. Control Signal Systems, 7 (1994), pp. 95–120.

[16] G.A. Leonov, I.M. Burkin, and A.I. Shepelyavyi, *Frequency Methods in Oscillation Theory*, Kluwer, Dordrecht, 1995 (in Russian: 1992).

[17] Y. Lin, E.D. Sontag, and Y.A. Wang, *Smooth converse Lyapunov theorem for robust stability*, SIAM J. Control Optim., 34 (1996), pp. 124–160.

[18] J. Mallet-Paret and G.R. Sell, *The Poincaré-Bendixson Theorem for monotone cyclic feedback systems with delay*, J. Differential Equations, 125 (1996), pp. 441–489.

[19] S. Martinez, J. Cortes, and F. Bullo, *Analysis and design of oscillatory control systems*, IEEE Trans. Automat. Control, 48 (2003), pp. 1164–1177.

[20] V.V. Nemytskii and V.V. Stepanov, *Qualitative Theory of Differential Equations*, Dover, New York, 1989.

[21] E.D. Sontag, *Smooth stabilization implies coprime factorization*, IEEE Trans. Automat. Control, 34 (1989), pp. 435–443.

[22] E.D. Sontag, *Asymptotic amplitudes and Cauchy gains: A small gain principle and an application to inhibitory biological feedback*, Systems Control Lett., 47 (2002), pp. 167–179.

[23] E.D. Sontag and Y. Wang, *Various results concerning set input-to-state stability*, Proc. IEEE CDC 95, IEEE Publications, 1995, pp. 1330–1335.

[24] E.D. Sontag and Y. Wang, *Notions of input to output stability*, Systems Control Lett., 38 (1999), pp. 235–248.

[25] E.D. Sontag and Y. Wang, *New characterization of the input to state stability property*, IEEE Trans. Automat. Control, 41 (1996), pp. 1283–1294.

[26] E.D. Sontag and Y. Wang, *Output-to-state stability and detectability of nonlinear systems*, Systems Control Lett., 29 (1997), pp. 279–290.

[27] R. Sepulchre, M. Janković, and P.V. Kokotović, *Constructive Nonlinear Control*, Springer-Verlag, New York, 1996.

[28] G.-B. Stan and R. Sepulchre, *Global analysis of limit cycles in networks of oscillators*, Proc. 5th IFAC Symposium on Nonlinear Control System (NOLCOS'04), 2004, Stuttgart, pp. 1433–1438.

[29] *Wikipedia, the Free Encyclopedia*, http://en.wikipedia.org/wiki/Oscillation_(mathematics), 2000.

[30] J.C. Willems, *Dissipative dynamical systems – part* I: *General theory*, Arch. Ration. Mech. Anal., 45 (1972), pp. 321–351.

[31] V.A. Yakubovich, *Frequency oscillations conditions in nonlinear systems with stationary single nonlinearity*, Siberian Math J., 14 (1973).

[32] V.A. Yakubovich, *Oscillations in systems with discontinuous and hysteresis nonlinearities*, Autom. Remote Control, 12 (1975).

[33] V.A. Yakubovich and E.A. Tomberg, *Conditions for self-induced oscillations in nonlinear systems*, Siberian Math. J., 30 (1989), pp. 641–653.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 2, pp. 641–671

© 2009 Society for Industrial and Applied Mathematics

APPROXIMATE TRACKING AND DISTURBANCE REJECTION FOR STABLE INFINITE-DIMENSIONAL SYSTEMS USING SAMPLED-DATA LOW-GAIN CONTROL*

ZHENQING KE†, HARTMUT LOGEMANN†, AND RICHARD REBARBER‡

Abstract. In this paper we solve tracking and disturbance rejection problems for stable infinite-dimensional systems using a simple low-gain controller suggested by the internal model principle. For stable discrete-time systems, it is shown that the application of a low-gain controller (depending on only one gain parameter) leads to a stable closed-loop system which asymptotically tracks reference signals r of the form $r(k) = \sum_{j=1}^{N} \lambda_j^k \mathfrak{r}_j$, where $\mathfrak{r}_j \in \mathbb{C}^p$ and $\lambda_j \in \mathbb{C}$ with $|\lambda_j| = 1$ for $j = 1, \dots, N$. The closed-loop system also rejects disturbance signals which are asymptotically of this form. The discrete-time result is used to derive results on approximate tracking and disturbance rejection for a large class of infinite-dimensional sampled-data feedback systems, with reference signals which are finite sums of sinusoids, and disturbance signals which are asymptotic to finite sums of sinusoids. The results are given for both input-output systems and state-space systems.

Key words. discrete-time systems, disturbance rejection, infinite-dimensional systems, internal model principle, low-gain control, sampled-data control, tracking

AMS subject classifications. 93C25, 93C55, 93C57, 93C80, 93D15, 93D25

DOI. 10.1137/080716517

1. Introduction.

The synthesis of low-gain integral controllers for uncertain stable continuous-time plants has received considerable attention in the last thirty years. Let $\mathbf{G}$ be a stable proper rational continuous-time transfer function matrix. The main existence result for robust low-gain integral control states that if all of the eigenvalues of $\mathbf{G}(0)$ have positive real parts, then there exists $\varepsilon^* > 0$ such that for all $\varepsilon \in (0, \varepsilon^*)$, the controller $(\varepsilon/s)I$ stabilizes $\mathbf{G}$. Moreover, the resulting closed-loop system asymptotically tracks arbitrary constant reference signals. This result has been proved by Davison [2] using state-space methods and Morari [11] using frequency-domain methods. This low-gain controller allows stabilization and tracking with very little information about the plant, and it is not based on system identification. The above regulator result has been extended to various classes of (abstract) infinite-dimensional continuous-time systems: in [12] for exponentially stable parabolic systems, in [7] for systems in the Callier–Desoer algebra (CD-algebra), and in [9] for exponentially stable regular systems.

In the case that the reference and disturbance signals are of the form

$$\sum_{j=1}^{N} e^{i\omega_j t}\mathfrak{w}_j\,, \quad \omega_j \in \mathbb{R}\,,\ \mathfrak{w}_j \in \mathbb{C}^m\,,$$

Hämäläinen and Pohjolainen [3] solved the tracking and disturbance rejection problem for stable infinite-dimensional systems in the CD-algebra. (In their paper, reference

*Received by the editors February 23, 2008; accepted for publication (in revised form) October 17, 2008; published electronically February 13, 2009. This work was supported by EPSRC grant GR/S94582/01, LMS Scheme 2 grant 2515, and NSF grant DMS-0606847.

http://www.siam.org/journals/sicon/48-2/71651.html

†Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK (kezhenqing@hotmail.com, hl@maths.bath.ac.uk).

‡Department of Mathematics, University of Nebraska, Lincoln, NE 68588-0130 (rrebarbe@math.unl.edu).

and disturbance signals are more general, containing polynomial parts.) Rebarber and Weiss [13] proved similar results for the more general class of exponentially stable well-posed systems.

In this paper, we consider low-gain control for infinite-dimensional discrete-time and sampled-data feedback systems. In section 2, we give preliminary technical results. In section 3, we develop a frequency-domain approach to discrete-time low-gain control. We consider a feedback controller of the form

$$\varepsilon\left(\mathbf{K}^0(z)+\sum_{j=1}^{N}\frac{K_j}{z-\lambda_j}\right), \tag{1.1}$$

where $\mathbf{K}^0$ has impulse response in $\ell^1(\mathbb{Z}_+,\mathbb{C}^{m\times p})$, $K_j\in\mathbb{C}^{m\times p}$, and $\lambda_j\in\mathbb{C}$ with $|\lambda_j|=1$. We assume that the plant has a transfer function $\mathbf{G}$ which has impulse response in $\ell^1(\mathbb{Z}_+,\mathbb{C}^{p\times m})$. We show that the application of this controller to the plant will result in an ℓ^q-stable closed-loop system for $1\le q\le\infty$, provided that

(i) all the eigenvalues of $\bar{\lambda}_j\mathbf{G}(\lambda_j)K_j$ have positive real parts;

(ii) $\limsup_{z\to\lambda_j\ |z|>1}\|(\mathbf{G}(z)-\mathbf{G}(\lambda_j))/(z-\lambda_j)\|<\infty$;

(iii) the gain parameter ε is sufficiently small.

Moreover, the closed-loop system achieves asymptotic tracking and disturbance rejection for reference signals r of the form $r(k):=\sum_{j=1}^N\lambda_j^k\mathfrak{r}_j$ and disturbance signals d satisfying $\lim_{k\to\infty}(d(k)-\sum_{j=1}^N\lambda_j^k\mathfrak{d}_j)=0$, where $\mathfrak{r}_j\in\mathbb{C}^p$ and $\mathfrak{d}_j\in\mathbb{C}^m$. The results are first proved for input-output systems, and then for state-space systems. Our results are an extension of results by Logemann and Townley [10]. In their paper, the reference and disturbance signals are constants.

In section 4, the discrete-time results in section 3 are used to derive results on approximate tracking and disturbance rejection for input-output and state-space sampled-data systems. The input-output operator G of the continuous-time plant is assumed to be a convolution operator of the form $Gu=\mu\star u$, where μ is a $\mathbb{C}^{p\times m}$-valued Borel measure such that $\int_{\mathbb{R}_+}e^{-\alpha t}|\mu|(dt)<\infty$ for some $\alpha<0$, where $|\mu|$ is the total variation of μ. The discrete-time controller underlying the sampled-data feedback scheme is given by (1.1) with $\lambda_j=e^{\xi_j\tau}$, where $\xi_j\in i\mathbb{R}$ for $j=1,\ldots,N$ and $\tau>0$ is the sampling period. The reference signals r are given by $r(t)=\sum_{j=1}^N e^{\xi_j t}\mathfrak{r}_j$, where $\mathfrak{r}_j\in\mathbb{C}^p$. Invoking both time-domain and frequency-domain methods, we prove that if all the eigenvalues of $\mathbf{G}(\xi_j)K_j$ have positive real parts, then, for every $\delta>0$, there exists $\tau_\delta>0$ such that, for every sampling period $\tau\in(0,\tau_\delta)$, there exists $\varepsilon_\tau>0$ such that, for every $\varepsilon\in(0,\varepsilon_\tau)$, the output y of the closed-loop sampled-data system satisfies

$$\limsup_{t\to\infty}\|y(t)-r(t)\|\le\delta$$

in the presence of disturbance signals d satisfying $\lim_{t\to\infty}(d(t)-\sum_{j=1}^N e^{\xi_j t}\mathfrak{d}_j)=0$, where $\mathfrak{d}_j\in\mathbb{C}^m$. At the end of the section we give an application to a heat equation.

To the best of our knowledge, the main results in sections 3 and 4 are new even for finite-dimensional systems.

Notation. Let X and Y be Banach spaces. The set of all bounded linear operators from X to Y is denoted by $\mathcal{B}(X,Y)$; we write $\mathcal{B}(X)$ for $\mathcal{B}(X,X)$. Moreover, $F(\mathbb{Z}_+,X)$ denotes all X-valued sequences defined on $\mathbb{Z}_+$, and $L_b(\mathbb{R}_+,X)$ denotes the set of bounded X-valued Lebesgue measurable functions with the sup-norm $\|\cdot\|_\infty$.

The z-transform of $v \in F(\mathbb{Z}_+, X)$ is denoted by $\mathscr{Z}(v)$. Sometimes we write $\hat{v}$ for $\mathscr{Z}(v)$. For $\alpha > 0$ and $\beta \in \mathbb{R}$, define $\mathbb{E}_\alpha := \{z \in \mathbb{C} : |z| > \alpha\}$ and $\mathbb{C}_\beta := \{z \in \mathbb{C} : \operatorname{Re} z > \beta\}$.

Let $\Omega \subset \mathbb{C}$ be open. We define

$$H^\infty(\Omega, \mathbb{C}^{p\times m}) := \{f : \Omega \to \mathbb{C}^{p\times m} \mid f \text{ is holomorphic and bounded}\},$$
$$H^\infty_<(\mathbb{E}_1, \mathbb{C}^{p\times m}) := \bigcup_{0<\gamma<1} H^\infty(\mathbb{E}_\gamma, \mathbb{C}^{p\times m}).$$

We write $H^\infty(\Omega) := H^\infty(\Omega, \mathbb{C})$. Let $\mathcal{Q}$ denote the quotient field of $H^\infty(\mathbb{E}_1)$, i.e., $\mathcal{Q} = \{n/d : n, d \in H^\infty(\mathbb{E}_1), d \neq 0\}$. Furthermore, let $\mathcal{R}_{\mathrm{s}}$ denote the ring of discrete-time stable proper complex rational functions, i.e., rational functions with complex coefficients which are bounded at ∞ and have all their poles in $\{z \in \mathbb{C} : |z| < 1\}$.

For $\alpha > 0$, define the weighted ℓ^1-space $\ell^1_\alpha(\mathbb{Z}_+, \mathbb{C}^{p\times m})$ by

$$\ell^1_\alpha(\mathbb{Z}_+, \mathbb{C}^{p\times m}) := \{v \in F(\mathbb{Z}_+, \mathbb{C}^{p\times m}) : (v(k)\alpha^{-k})_{k\in\mathbb{Z}_+} \in \ell^1(\mathbb{Z}_+, \mathbb{C}^{p\times m})\}$$

and set

$$\hat{\ell}^1_\alpha(\mathbb{C}^{p\times m}) := \{\mathscr{Z}(g) : g \in \ell^1_\alpha(\mathbb{Z}_+, \mathbb{C}^{p\times m})\} \subset H^\infty(\mathbb{E}_\alpha, \mathbb{C}^{p\times m}).$$

We write $\hat{\ell}^1(\mathbb{C}^{p\times m}) := \hat{\ell}^1_1(\mathbb{C}^{p\times m})$. For $A \in \mathcal{B}(X)$, let $\sigma(A)$ denote the spectrum of A. For $N \in \mathbb{N}$, set $\underline{N} := \{1, 2, \ldots, N\}$. Finally, throughout, the symbol $\star$ denotes convolution (in discrete and continuous time).

2. Preliminaries. Let $\mathcal{F}(\mathbf{G}, \mathbf{K})$ denote the (discrete-time) feedback system shown in Figure 2.1, where $\mathbf{G} \in \mathcal{Q}^{p\times m}$ and $\mathbf{K} \in \mathcal{Q}^{m\times p}$. For $(\mathbf{G}, \mathbf{K}) \in \mathcal{Q}^{p\times m} \times \mathcal{Q}^{m\times p}$ such that $\det(I + \mathbf{GK}) \neq 0$, we set

$$F(\mathbf{G}, \mathbf{K}) := \begin{pmatrix} (I + \mathbf{GK})^{-1} & \mathbf{G}(I + \mathbf{KG})^{-1} \\ \mathbf{K}(I + \mathbf{GK})^{-1} & (I + \mathbf{KG})^{-1} \end{pmatrix}. \tag{2.1}$$

The feedback system $\mathcal{F}(\mathbf{G}, \mathbf{K})$ is called *ℓ^q-stable* (where $1 \le q \le \infty$) if there exists $M \ge 0$ such that, for all r, $d_2 \in \ell^q(\mathbb{Z}_+, \mathbb{C}^p)$ and all $d_1 \in \ell^q(\mathbb{Z}_+, \mathbb{C}^m)$,

$$\|y_p\|_{\ell^q} + \|y_c\|_{\ell^q} \le M(\|r\|_{\ell^q} + \|d_1\|_{\ell^q} + \|d_2\|_{\ell^q}).$$

It is easy to see that $\mathcal{F}(\mathbf{G}, \mathbf{K})$ is ℓ^q-stable if $F(\mathbf{G}, \mathbf{K}) \in \hat{\ell}^1(\mathbb{C}^{(m+p)\times(m+p)})$, and it is a standard result that $\mathcal{F}(\mathbf{G}, \mathbf{K})$ is ℓ^2-stable if and only if $F(\mathbf{G}, \mathbf{K}) \in H^\infty(\mathbb{E}_1, \mathbb{C}^{(m+p)\times(m+p)})$.

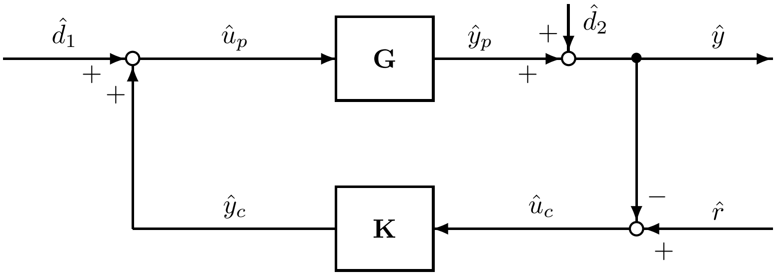


FIG. 2.1. *Discrete-time closed-loop system $\mathcal{F}(\mathbf{G}, \mathbf{K})$.*

Definition 2.1. *A* left-coprime factorization *of* $\mathbf{G} \in \mathcal{Q}^{p\times m}$ *(over* $H^\infty(\mathbb{E}_1)$*) is a pair* $(\mathbf{D}, \mathbf{N}) \in H^\infty(\mathbb{E}_1, \mathbb{C}^{p\times p}) \times H^\infty(\mathbb{E}_1, \mathbb{C}^{p\times m})$ *such that* $\det \mathbf{D} \neq 0$*,* $\mathbf{G} = \mathbf{D}^{-1}\mathbf{N}$ *and* $\mathbf{D}$ *and* $\mathbf{N}$ *are left coprime; i.e., there exist* $\mathbf{X} \in H^\infty(\mathbb{E}_1, \mathbb{C}^{p\times p})$*,* $\mathbf{Y} \in H^\infty(\mathbb{E}_1, \mathbb{C}^{m\times p})$ *satisfying* $\mathbf{DX} + \mathbf{NY} = I$*.*

A right-coprime factorization *of* $\mathbf{G} \in \mathcal{Q}^{p\times m}$ *(over* $H^\infty(\mathbb{E}_1)$*) is a pair* $(\mathbf{N}, \mathbf{D}) \in H^\infty(\mathbb{E}_1, \mathbb{C}^{p\times m}) \times H^\infty(\mathbb{E}_1, \mathbb{C}^{m\times m})$ *such that* $\det \mathbf{D} \neq 0$*,* $\mathbf{G} = \mathbf{N}\mathbf{D}^{-1}$ *and* $\mathbf{N}$ *and* $\mathbf{D}$ *are right coprime; i.e., there exist* $\mathbf{X} \in H^\infty(\mathbb{E}_1, \mathbb{C}^{m\times p})$*,* $\mathbf{Y} \in H^\infty(\mathbb{E}_1, \mathbb{C}^{m\times m})$ *satisfying* $\mathbf{XN} + \mathbf{YD} = I$*.*

Remark 2.2. It follows from [14] that $\mathbf{G}$ and $\mathbf{K}$ admit left- and right-coprime factorizations (over $H^\infty(\mathbb{E}_1)$) if $\mathcal{F}(\mathbf{G}, \mathbf{K})$ is ℓ^2-stable.

An application of a standard result in fractional representation theory (see [17, Lemma 3.1]) gives the following necessary and sufficient algebraic condition for closed-loop stability in terms of coprime factors.

Proposition 2.3. *Let* $\mathbf{G} \in \mathcal{Q}^{p\times m}$ *and* $\mathbf{K} \in \mathcal{Q}^{m\times p}$*. Assume that there exist a left-coprime factorization* $(\mathbf{D_G}, \mathbf{N_G})$ *of* $\mathbf{G}$ *and a right-coprime factorization* $(\mathbf{N_K}, \mathbf{D_K})$ *of* $\mathbf{K}$ *(both over* $H^\infty(\mathbb{E}_1)$*). Then the feedback system* $\mathcal{F}(\mathbf{G}, \mathbf{K})$ *is* ℓ^2*-stable if and only if the matrix* $\mathbf{N_G N_K} + \mathbf{D_G D_K}$ *has an inverse in* $H^\infty(\mathbb{E}_1, \mathbb{C}^{p\times p})$*, i.e., if and only if*

$$\inf_{z\in\mathbb{E}_1} |\det[\mathbf{N_G}(z)\mathbf{N_K}(z) + \mathbf{D_G}(z)\mathbf{D_K}(z)]| > 0\,.$$

Proposition 2.4 (see [1, Lemma 3.1]). *Assume that* $\mathbf{G} \in \hat{\ell}^1(\mathbb{C}^{m\times m})$*. Then* $\mathbf{G}$ *has an inverse in* $\hat{\ell}^1(\mathbb{C}^{m\times m})$ *if and only if*

$$\inf_{z\in\mathbb{E}_1} |\det \mathbf{G}(z)| > 0\,.$$

The next result will be an important tool in the proof of our main theorem in section 3, and it is also interesting in its own right.

Proposition 2.5. *Let* $\mathbf{G} \in \mathcal{Q}^{p\times m}$ *and* $\mathbf{K} \in \mathcal{Q}^{m\times p}$*. Assume that the feedback system* $\mathcal{F}(\mathbf{G}, \mathbf{K})$ *is* ℓ^2*-stable. Let* $(\mathbf{D_G}, \mathbf{N_G})$ *be a left-coprime factorization of* $\mathbf{G}$ *and* $(\mathbf{N_K}, \mathbf{D_K})$ *be a right-coprime factorization of* $\mathbf{K}$ *(both over* $H^\infty(\mathbb{E}_1)$*). Assume that* $\mathbf{D_G}, \mathbf{D_K} \in \hat{\ell}^1(\mathbb{C}^{p\times p})$*,* $\mathbf{N_G} \in \hat{\ell}^1(\mathbb{C}^{p\times m})$*, and* $\mathbf{N_K} \in \hat{\ell}^1(\mathbb{C}^{m\times p})$*. Then* $F(\mathbf{G}, \mathbf{K}) \in \hat{\ell}^1(\mathbb{C}^{(m+p)\times(m+p)})$*. In particular,* $\mathcal{F}(\mathbf{G}, \mathbf{K})$ *is* ℓ^q*-stable for* $1 \le q \le \infty$*.*

Proof. By hypothesis, it is clear that $\mathbf{N_G N_K}+\mathbf{D_G D_K} \in \hat{\ell}^1(\mathbb{C}^{p\times p})$. Since $\mathcal{F}(\mathbf{G}, \mathbf{K})$ is ℓ^2-stable, by Proposition 2.3,

$$\inf_{z\in\mathbb{E}_1} |\det[\mathbf{N_G}(z)\mathbf{N_K}(z) + \mathbf{D_G}(z)\mathbf{D_K}(z)]| > 0\,.$$

Then it follows from Proposition 2.4 that $(\mathbf{N_G N_K} + \mathbf{D_G D_K})^{-1} \in \hat{\ell}^1(\mathbb{C}^{p\times p})$. It is easy to see that

$$(I + \mathbf{GK})^{-1} = \mathbf{D_K}(\mathbf{N_G N_K} + \mathbf{D_G D_K})^{-1}\mathbf{D_G}\,,$$

so that $(I + \mathbf{GK})^{-1} \in \hat{\ell}^1(\mathbb{C}^{p\times p})$. By simple calculations, we obtain

$$\begin{aligned}
\mathbf{K}(I + \mathbf{GK})^{-1} &= \mathbf{N_K}(\mathbf{N_G N_K} + \mathbf{D_G D_K})^{-1}\mathbf{D_G}\,,\\
\mathbf{G}(I + \mathbf{KG})^{-1} &= (I + \mathbf{GK})^{-1}\mathbf{G} = \mathbf{D_K}(\mathbf{N_G N_K} + \mathbf{D_G D_K})^{-1}\mathbf{N_G}\,,\\
(I + \mathbf{KG})^{-1} &= I - \mathbf{K}(I + \mathbf{GK})^{-1}\mathbf{G} = I - \mathbf{N_K}(\mathbf{N_G N_K} + \mathbf{D_G D_K})^{-1}\mathbf{N_G}\,,
\end{aligned}$$

showing that $\mathbf{K}(I + \mathbf{GK})^{-1}$, $\mathbf{G}(I + \mathbf{KG})^{-1}$, and $(I + \mathbf{KG})^{-1}$ have all their entries in $\hat{\ell}^1(\mathbb{C})$. Hence $F(\mathbf{G}, \mathbf{K}) \in \hat{\ell}^1(\mathbb{C}^{(m+p)\times(m+p)})$. □

The following frequency-response result for transfer functions in $\hat{\ell}^1(\mathbb{C}^{p\times m})$ will be useful for understanding the asymptotic behavior of the closed-loop system.

LEMMA 2.6. *Let* $g \in F(\mathbb{Z}_+, \mathbb{C}^{p\times m})$, $u \in F(\mathbb{Z}_+, \mathbb{C}^m)$, $\lambda \in \overline{\mathbb{E}}_1$, $\mathfrak{v} \in \mathbb{C}^m$ *and set* $\mathbf{G} := \mathscr{Z}(g)$.

1. *If* $g \in \ell^1(\mathbb{Z}_+, \mathbb{C}^{p\times m})$ *and* $\lim_{n\to\infty}(u(n) - \lambda^n\mathfrak{v}) = 0$, *then*

$$\lim_{n\to\infty}[(g \star u)(n) - \lambda^n\mathbf{G}(\lambda)\mathfrak{v}] = 0\,.$$

2. *If there exist* $\beta \in (0,1)$ *and* $M \geq 0$ *such that* $g \in \ell^1_\beta(\mathbb{Z}_+, \mathbb{C}^{p\times m})$ *and*

$$\|u(n) - \lambda^n\mathfrak{v}\| \leq M\beta^n \quad \forall n \in \mathbb{Z}_+\,,$$

then there exists $L \geq 0$ *such that*

$$\|(g \star u)(n) - \mathbf{G}(\lambda)\lambda^n\mathfrak{v}\| \leq L\beta^n \quad \forall n \in \mathbb{Z}_+\,.$$

Proof. Since $g \in \ell^1(\mathbb{Z}_+, \mathbb{C}^{p\times m})$,

$$\|\mathbf{G}(z)\| = \left\|\sum_{k=0}^{\infty} g(k)z^{-k}\right\| \leq \sum_{k=0}^{\infty}\|g(k)\||z|^{-k} \leq \sum_{k=0}^{\infty}\|g(k)\| < \infty \quad \forall z \in \overline{\mathbb{E}}_1\,,$$

so that $\mathbf{G}(z)$ is well defined for $z \in \overline{\mathbb{E}}_1$. Define $v \in F(\mathbb{Z}_+, \mathbb{C}^m)$ by $v(k) := \lambda^k\mathfrak{v}$. Since $\lambda \in \overline{\mathbb{E}}_1$, $|\lambda|^{-k} \leq 1$ for all $k \in \mathbb{Z}_+$. Therefore,

$$\begin{aligned}
\|(g \star u)(n) - \lambda^n\mathbf{G}(\lambda)\mathfrak{v}\| &= \left\|\sum_{k=0}^{n} g(k)u(n-k) - \sum_{k=0}^{\infty}\lambda^{n-k}g(k)\mathfrak{v}\right\| \\
&\leq \left\|\sum_{k=0}^{n} g(k)(u(n-k) - v(n-k))\right\| + \|\mathfrak{v}\|\sum_{k=n+1}^{\infty}|\lambda|^{n-k}\|g(k)\| \\
&\leq \|(g \star (u-v))(n)\| + \|\mathfrak{v}\|\sum_{k=n}^{\infty}\|g(k)\| \quad \forall n \in \mathbb{Z}_+\,. \qquad (2.2)
\end{aligned}$$

We proceed to prove statement 1. Let $M_1 \geq 0$ be such that $\|u(k) - v(k)\| \leq M_1$ for all $k \in \mathbb{Z}_+$. By hypothesis, $\lim_{k\to\infty}\|u(k) - v(k)\| = 0$ and $g \in \ell^1(\mathbb{Z}_+, \mathbb{C}^{p\times m})$. Therefore, for $\varepsilon > 0$, there exists $k_0 \in \mathbb{Z}_+$ such that

$$\|u(k) - v(k)\| \leq \frac{\varepsilon}{2\|g\|_{\ell^1}}\,, \quad \sum_{j=k}^{\infty}\|g(j)\| \leq \frac{\varepsilon}{2M_1}\,; \quad \forall k \geq k_0.$$

Then, for $n \geq 2k_0$,

$$\begin{aligned}
\|(g \star (u-v))(n)\| &\leq \sum_{k=0}^{k_0}\|g(k)\|\|(u-v)(n-k)\| + \sum_{k=k_0+1}^{n}\|g(k)\|\|(u-v)(n-k)\| \\
&\leq \frac{\varepsilon}{2\|g\|_{\ell^1}}\sum_{k=0}^{k_0}\|g(k)\| + M_1\sum_{k=k_0+1}^{n}\|g(k)\| \\
&\leq \varepsilon\,,
\end{aligned}$$

showing that

$$\lim_{n\to\infty} \|g \star (u-v)(n)\| = 0\,. \tag{2.3}$$

A combination of (2.2), (2.3), and the fact that $\lim_{n\to\infty}\sum_{k=n}^{\infty}\|g(k)\| = 0$ yields statement 1.

To prove statement 2, we set $M_2 := \sum_{k=0}^{\infty}\beta^{-k}\|g(k)\| < \infty$. By hypothesis, there exists $M \geq 0$ such that

$$\|(u-v)(n)\| \leq M\beta^n \quad \forall n \in \mathbb{Z}_+\,.$$

Since $\beta \in (0,1)$ and by (2.2), we have

$$\begin{aligned}\beta^{-n}\|(g\star u)(n) - \mathbf{G}(\lambda)\lambda^n\mathfrak{v}\| &\leq \beta^{-n}\sum_{k=0}^{n}\|g(k)\|\|(u-v)(n-k)\| + \beta^{-n}\|\mathfrak{v}\|\sum_{k=n}^{\infty}\|g(k)\| \\ &\leq \beta^{-n}\sum_{k=0}^{n}\|g(k)\|M\beta^{n-k} + \|\mathfrak{v}\|\sum_{k=n}^{\infty}\beta^{-k}\|g(k)\| \\ &\leq MM_2 + \|\mathfrak{v}\|M_2 \quad \forall n \in \mathbb{Z}_+\,.\end{aligned}$$

Hence $\|(g\star u)(n) - \mathbf{G}(\lambda)\lambda^n\mathfrak{v}\| \leq M_2(M + \|\mathfrak{v}\|)\beta^n$ for all $n \in \mathbb{Z}_+$. □

The next result shows that Lemma 2.6 applies in particular to input-output operators with transfer functions in $H^{\infty}_{<}(\mathbb{E}_1, \mathbb{C}^{p\times m})$. We omit the routine proof.

Proposition 2.7. *For* $0 < \alpha < \beta$*,* $H^{\infty}(\mathbb{E}_\alpha, \mathbb{C}^{p\times m}) \subset \hat{\ell}^1_\beta(\mathbb{C}^{p\times m})$*.*

The following remark shows that Lemma 2.6 also applies to power stable state-space systems.

Remark 2.8. Consider a discrete-time state-space system

$$x_p(k+1) = Ax_p(k) + Bu_p(k)\,, \tag{2.4a}$$

$$y_p(k) = Cx_p(k) + Du_p(k)\,, \tag{2.4b}$$

evolving on a Banach space X, where $A \in \mathcal{B}(X)$, $B \in \mathcal{B}(\mathbb{C}^m, X)$, $C \in \mathcal{B}(X, \mathbb{C}^p)$, and $D \in \mathcal{B}(\mathbb{C}^m, \mathbb{C}^p)$. The transfer function $\mathbf{G}$ of (2.4) is given by

$$\mathbf{G}(z) = C(zI - A)^{-1}B + D\,.$$

System (2.4) is called *power stable* if A is power stable, i.e., there exist $M \geq 1$ and $\rho \in (0,1)$ such that

$$\|A^k\| \leq M\rho^k \quad \forall k \in \mathbb{Z}_+\,.$$

Clearly, if (2.4) is power stable, then $\sigma(A) \subset \{z \in \mathbb{C} : |z| < 1\}$ and $\mathbf{G} \in H^{\infty}_{<}(\mathbb{E}_1, \mathbb{C}^{p\times m})$. Hence, by Proposition 2.7, Lemma 2.6 applies to power stable systems of the form (2.4).

3. Low-gain control of discrete-time systems. Let $\mathcal{F}(\mathbf{G}, \mathbf{K}_\varepsilon)$ denote the discrete-time feedback system shown in Figure 2.1 and given by (2.1), with $\mathbf{K}$ replaced with $\mathbf{K}_\varepsilon$. The following asymptotic tracking theorem is the main result of this section. It is the discrete-time counterpart of the continuous-time result due to Rebarber and Weiss [13], which is a partial extension of the main results in Hämäläinen and Pohjolainen [3].

THEOREM 3.1. *Let $N \in \mathbb{N}$. For $j \in \underline{N}$, let $\lambda_j \in \mathbb{C}$ be such that $|\lambda_j| = 1$ and $\lambda_j \neq \lambda_k$ for $j \neq k$. Assume that $\mathbf{G} \in \hat{\ell}^1(\mathbb{C}^{p\times m})$ and $\mathbf{K}_\varepsilon$ is given by*

$$\mathbf{K}_\varepsilon(z) := \varepsilon \left(\mathbf{K}^0(z) + \sum_{j=1}^{N} \frac{K_j}{z - \lambda_j} \right) , \tag{3.1}$$

where $\mathbf{K}^0 \in \hat{\ell}^1(\mathbb{C}^{m\times p})$ and $K_j \in \mathbb{C}^{m\times p}$. If

$$\sigma[\bar{\lambda}_j \mathbf{G}(\lambda_j) K_j] \subset \mathbb{C}_0 \quad \forall j \in \underline{N} \tag{3.2}$$

and

$$\limsup_{z\to\lambda_j,\, z\in\mathbb{E}_1} \left\| \frac{\mathbf{G}(z) - \mathbf{G}(\lambda_j)}{z - \lambda_j} \right\| < \infty \quad \forall j \in \underline{N}, \tag{3.3}$$

then there exists $\varepsilon^* > 0$ such that, for all $\varepsilon \in (0, \varepsilon^*)$, we have $F(\mathbf{G}, \mathbf{K}_\varepsilon) \in \hat{\ell}^1(\mathbb{C}^{(m+p)\times(m+p)})$ (thus $\mathcal{F}(\mathbf{G}, \mathbf{K}_\varepsilon)$ is ℓ^q-stable for every $1 \leq q \leq \infty$).

Moreover, if the reference signal r is given by

$$r(k) := \sum_{j=1}^{N} \lambda_j^k \mathfrak{r}_j \, , \ \mathfrak{r}_j \in \mathbb{C}^p \, , \quad \forall k \in \mathbb{Z}_+ \, , \tag{3.4}$$

and the disturbance signals d_1, d_2 satisfy

$$\lim_{k\to\infty} \left(d_1(k) - \sum_{j=1}^{N} \lambda_j^k \mathfrak{d}_{1j} \right) = 0 \, , \quad \lim_{k\to\infty} \left(d_2(k) - \sum_{j=1}^{N} \lambda_j^k \mathfrak{d}_{2j} \right) = 0 \tag{3.5}$$

$$\textit{for some } \mathfrak{d}_{1j} \in \mathbb{C}^m \textit{ and } \mathfrak{d}_{2j} \in \mathbb{C}^p,$$

then, for every $\varepsilon \in (0, \varepsilon^*)$, the output of the closed-loop system y asymptotically tracks r in the presence of d_1, d_2, that is, $\lim_{k\to\infty}(y(k) - r(k)) = 0$.

Remark 3.2. (i) If condition (3.2) does not hold, then there is no guarantee that there exists an $\varepsilon > 0$ such that the feedback system $\mathcal{F}(\mathbf{G}, \mathbf{K}_\varepsilon)$ is ℓ^2-stable. Indeed, if $N = m = p = 1$, $\lambda_1 = 1$, $K_1 = 1$, and $\mathbf{G} \in \hat{\ell}^1(\mathbb{C})$ with $\mathbf{G}(1) \in (-\infty, 0]$, then an application of Proposition 2.3 shows that $\mathcal{F}(\mathbf{G}, \mathbf{K}_\varepsilon)$ is not ℓ^2-stable for every $\varepsilon > 0$. Furthermore, if $N = 1$ and $\lambda_1 = 1$, then it can be shown that the existence of an $\varepsilon^* > 0$ such that, for all $\varepsilon \in (0, \varepsilon^*)$, $\mathcal{F}(\mathbf{G}, \mathbf{K}_\varepsilon)$ is ℓ^2-stable implies that $\sigma(\mathbf{G}(1)K_1) \subset \overline{\mathbb{C}_0}$ (this follows from a suitable modification of an argument used in [11, Theorem 3]). Consequently, at least in the case $N = 1$ and $\lambda_1 = 1$, condition (3.2) is "close" to being necessary for the stability conclusion of Theorem 3.1 to hold.

(ii) Condition (3.3) is not very restrictive. It is, for example, satisfied if, for every $j \in \underline{N}$, the transfer function $\mathbf{G}$ has a holomorphic extension to an open neighborhood of λ_j (which is trivially the case if $\mathbf{G} \in H^\infty_<(\mathbb{E}_1, \mathbb{C}^{p\times m})$).

(iii) Note that only very little plant information is required in order to apply Theorem 3.1, namely, stability of the system to be controlled, condition (3.3), and some information on $\mathbf{G}(\lambda_j)$, where the latter is required for the computation of K_j such that (3.2) holds. The spectral condition (3.2) is robust with respect to "sufficiently small" plant perturbations, while (3.3) is robust with respect to all plant perturbation in $H^\infty_<(\mathbb{E}_1, \mathbb{C}^{p\times m})$.

(iv) If, in Theorem 3.1, we replace the controller $\mathbf{K}_\varepsilon$ by

$$\tilde{\mathbf{K}}_\varepsilon(z) := \varepsilon\left(\tilde{\mathbf{K}}^0(z) + \sum_{j=1}^{N}\frac{z\tilde{K}_j}{z-\lambda_j}\right),$$

where $\tilde{\mathbf{K}}^0 \in \hat{\ell}^1(\mathbb{C}^{m\times p})$ and $\tilde{K}_j \in \mathbb{C}^{m\times p}$, and condition (3.2) by

$$\sigma(\mathbf{G}(\lambda_j)\tilde{K}_j) \subset \mathbb{C}_0 \quad \forall j \in \underline{N}\,, \tag{3.6}$$

while all the other conditions in the theorem remain the same, then the conclusions on stability, tracking, and disturbance rejection in Theorem 3.1 are still valid. This follows directly from Theorem 3.1, since

$$\tilde{\mathbf{K}}_\varepsilon(z) = \varepsilon\left(\tilde{\mathbf{K}}^0(z) + \sum_{j=1}^{N}\tilde{K}_j + \sum_{j=1}^{N}\frac{\lambda_j\tilde{K}_j}{z-\lambda_j}\right)$$

is of the form (3.1) with

$$\mathbf{K}^0(z) = \tilde{\mathbf{K}}^0(z) + \sum_{j=1}^{N}\tilde{K}_j\,, \quad K_j = \lambda_j\tilde{K}_j\,,$$

and $\sigma(\bar{\lambda}_j\mathbf{G}(\lambda_j)K_j) = \sigma(\mathbf{G}(\lambda_j)\tilde{K}_j) \subset \mathbb{C}_0$.

(v) The spectral condition (3.2) (or, alternatively, (3.6)) is the discrete-time analogue of the continuous-time condition in [13]; see [13, equation (1.5)]. Moreover, in the continuous-time result [13, Theorem 1.1], it is assumed that the transfer function of the plant is holomorphic and bounded in a half-plane of the form $\operatorname{Re} s > -\alpha$ for some $\alpha > 0$; the discrete-time analogue of this condition is, in the terminology of the present paper, $\mathbf{G} \in H^\infty_<(\mathbb{E}_1, \mathbb{C}^{p\times m})$, which implies (3.3) (cf. part (ii) of this remark). Consequently, condition (3.3) is weaker than the corresponding continuous-time condition in [13, Theorem 1.1].

To facilitate the proof of Theorem 3.1, we first state and prove the following key lemma, which shows that the transfer function $(I+\mathbf{G}\mathbf{K}_\varepsilon)^{-1}$, the so-called sensitivity function, is in $H^\infty(\mathbb{E}_1, \mathbb{C}^{p\times p})$ for sufficiently small $\varepsilon > 0$.

LEMMA 3.3. *Let $N \in \mathbb{N}$ and let $\lambda_j \in \mathbb{C}$ be such that $|\lambda_j| = 1$ and $\lambda_j \neq \lambda_k$ for $j,k \in \underline{N}$, $j \neq k$. Let $\mathbf{G} \in H^\infty(\mathbb{E}_1, \mathbb{C}^{p\times m})$ be such that the limit $\mathbf{G}(\lambda_j) := \lim_{z\to\lambda_j\,,\, z\in\mathbb{E}_1}\mathbf{G}(z)$ exists for every $j \in \underline{N}$. Let $\mathbf{K}_\varepsilon$ be given by* (3.1)*, where $\mathbf{K}^0 \in H^\infty(\mathbb{E}_1, \mathbb{C}^{m\times p})$ and $K_j \in \mathbb{C}^{m\times p}$. Assume that* (3.2) *and* (3.3) *hold. Then there exists $\varepsilon^* > 0$ such that, for all $\varepsilon \in (0, \varepsilon^*)$, $(I+\mathbf{G}\mathbf{K}_\varepsilon)^{-1} \in H^\infty(\mathbb{E}_1, \mathbb{C}^{p\times p})$. Moreover, if the additional assumptions that $\mathbf{G} \in H^\infty_<(\mathbb{E}_1, \mathbb{C}^{p\times m})$ and $\mathbf{K}^0 \in H^\infty_<(\mathbb{E}_1, \mathbb{C}^{m\times p})$ are satisfied, then, for every $\varepsilon \in (0, \varepsilon^*)$, $(I+\mathbf{G}\mathbf{K}_\varepsilon)^{-1} \in H^\infty_<(\mathbb{E}_1, \mathbb{C}^{p\times p})$.*

Proof. Before proceeding to the technical details, we summarize the idea of the proof. We wish to show that $(I+\mathbf{G}\mathbf{K}_\varepsilon)^{-1}$ is bounded in $\mathbb{E}_1$, the complement of the closed unit disc, for all sufficiently small $\varepsilon > 0$. Roughly speaking, we decompose $\mathbb{E}_1$ in the form $\mathbb{E}_1 = \Omega \cup \left(\cup_{j=1}^{n}\Omega_j\right)$, where Ω is bounded away from all of the λ_j's and Ω_j is the "part of $\mathbb{E}_1$ near λ_j." We prove that, for sufficiently small $\varepsilon > 0$, $(I+\mathbf{G}\mathbf{K}_\varepsilon)^{-1}$ is bounded on each of these sets. Special care is required for the analysis on the sets Ω_j.

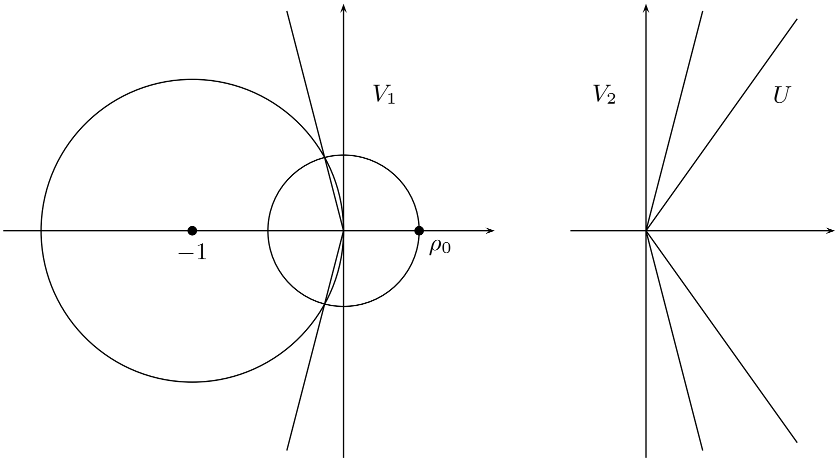


FIG. 3.1. *An illustration of the sets U, V_1, and V_2.*

Returning to the technical details of the proof, we first note that, since $\sigma[\bar{\lambda}_j\mathbf{G}(\lambda_j)K_j] \subset \mathbb{C}_0$ for all $j \in \underline{N}$, there exists $\theta \in (0, \pi/2)$ such that

$$\bigcup_{j=1}^{N} \sigma[\bar{\lambda}_j\mathbf{G}(\lambda_j)K_j] \subset \{z \in \mathbb{C} \setminus \{0\} : \arg z \in (-\theta, \theta)\} =: U\,. \tag{3.7}$$

Let $\rho \in (0, 1)$ and consider Figure 3.1. The circles $\{z \in \mathbb{C} : |z| = \rho\}$ and $\{z \in \mathbb{C} : |z + 1| = 1\}$ intersect at two points, denoted by $\rho e^{i\phi(\rho)}$ and $\rho e^{-i\phi(\rho)}$, where $\phi(\rho) \in (\pi/2, \pi)$. Note that $\phi(\rho) \to \pi/2$ monotonically as $\rho \to 0$. Hence there exists $\rho_0 \in (0, 1)$ such that $\pi - \phi(\rho) > \theta$ for all $\rho \in (0, \rho_0]$. Set

$$V_1 := \{z \in \mathbb{C} \setminus \{0\} : \arg z \in (-\phi(\rho_0), \phi(\rho_0))\}$$

and

$$V_2 := -V_1 = \{z \in \mathbb{C} \setminus \{0\} : \arg z \in (\pi - \phi(\rho_0), \pi + \phi(\rho_0))\}.$$

Clearly,

$$U \cap \overline{V}_2 = \emptyset\,. \tag{3.8}$$

There exists $\rho_1 \in (0, \rho_0]$ such that $|\lambda_j - \lambda_k| > 2\rho_1$ for all $j, k \in \underline{N}$, $j \neq k$. Defining

$$\Omega_j := \mathbb{E}_1 \bigcap \{z \in \mathbb{C} : |z - \lambda_j| < \rho_1\}\,,$$

we have that $\Omega_j \cap \Omega_k = \emptyset$ for $j, k \in \underline{N}$, $j \neq k$. Moreover, set $\Omega := \mathbb{E}_1 \setminus \bigcup_{j=1}^{N} \Omega_j$. Assume that $\mathbf{G} \in H^\infty(\mathbb{E}_1, \mathbb{C}^{p\times m})$ and $\mathbf{K}^0 \in H^\infty(\mathbb{E}_1, \mathbb{C}^{m\times p})$. It is clear that

$$\sup_{z\in\Omega} \left\| \mathbf{G}(z) \left(\mathbf{K}^0(z) + \sum_{j=1}^{N} \frac{K_j}{z - \lambda_j} \right) \right\| < \infty\,.$$

Therefore, there exists $\varepsilon_\infty > 0$ such that

$$\mathbf{S}(z) := [I + \mathbf{G}(z)\mathbf{K}_\varepsilon(z)]^{-1} = \left[I + \varepsilon\mathbf{G}(z) \left(\mathbf{K}^0(z) + \sum_{j=1}^{N} \frac{K_j}{z - \lambda_j} \right) \right]^{-1}$$

is uniformly bounded for all $z \in \Omega$ and for all $\varepsilon \in (0, \varepsilon_\infty)$. Fix $j \in \underline{N}$. To analyze $\mathbf{S}$ on Ω_j, we define

$$\mathbf{S}_j(z) := \left(I + \frac{\varepsilon \mathbf{G}(\lambda_j) K_j}{z - \lambda_j}\right)^{-1} = \left(I + \frac{\varepsilon \bar{\lambda}_j \mathbf{G}(\lambda_j) K_j}{\bar{\lambda}_j z - 1}\right)^{-1}$$

and

$$\mathbf{Q}_j(z) := \frac{\mathbf{G}(z) - \mathbf{G}(\lambda_j)}{z - \lambda_j} K_j + \mathbf{G}(z)\mathbf{K}^0(z) + \sum_{k \in \underline{N},\, k \neq j} \frac{\mathbf{G}(z) K_k}{z - \lambda_k}\,.$$

By (3.3), we see that $\mathbf{Q}_j$ is bounded on Ω_j, with a bound that is independent of ε. For convenience, we set $G_j := \bar{\lambda}_j \mathbf{G}(\lambda_j) K_j$. Moreover, since $\rho_1 \in (0, \rho_0]$, it follows that $\bar{\lambda}_j \Omega_j - 1 \subset V_1$. Together with the implication that if $w \in V_1$, then $\gamma w \in V_1$ for all $\gamma \geq 0$, this yields

$$\begin{aligned}\sup_{z \in \Omega_j} \|\mathbf{S}_j(z)\| &= \sup \left\{ \left\| \left(I + \varepsilon \frac{G_j}{w}\right)^{-1} \right\| : w \in \bar{\lambda}_j \Omega_j - 1 \right\} \\ &\leq \sup_{s \in V_1} \|s(sI + G_j)^{-1}\| = \sup_{s \in V_2} \|s(sI - G_j)^{-1}\|\,.\end{aligned}$$

By (3.7) and (3.8), the function $s \mapsto s(sI - G_j)^{-1}$ is holomorphic on an open set $W \supset \overline{V}_2$ (where $\overline{V}_2$ denotes the closure of V_2). Furthermore,

$$\lim_{|s| \to \infty} s(sI - G_j)^{-1} = I\,.$$

Hence $s \mapsto s(sI - G_j)^{-1}$ is bounded on $\overline{V}_2$. Therefore, $\mathbf{S}_j$ is bounded on Ω_j with bound independent of ε. We have $\mathbf{S}^{-1} - \mathbf{S}_j^{-1} = \varepsilon \mathbf{Q}_j$, so that we can write

$$\mathbf{S}(z) = \mathbf{S}_j(z)(I + \varepsilon \mathbf{Q}_j(z) \mathbf{S}_j(z))^{-1}\,.$$

Hence there exists $\varepsilon_j \in (0, \varepsilon_\infty)$ such that $\mathbf{S}$ is bounded on Ω_j for all $\varepsilon \in (0, \varepsilon_j)$. Setting

$$\varepsilon^* := \min\{\varepsilon_j : j \in \underline{N}\}\,,$$

it follows that

$$(I + \mathbf{G}\mathbf{K}_\varepsilon)^{-1} \in H^\infty(\mathbb{E}_1, \mathbb{C}^{p \times p}) \quad \forall \varepsilon \in (0, \varepsilon^*)\,. \tag{3.9}$$

Finally, let $\varepsilon \in (0, \varepsilon^*)$ and assume that $\mathbf{G} \in H^\infty_<(\mathbb{E}_1, \mathbb{C}^{p \times m})$ and $\mathbf{K}^0 \in H^\infty_<(\mathbb{E}_1, \mathbb{C}^{m \times p})$. It is clear that $(I + \mathbf{G}\mathbf{K}_\varepsilon)^{-1}$ is meromorphic on $\mathbb{E}_\gamma$ for some $\gamma \in (0, 1)$. Letting $\beta \in (\gamma, 1)$, it follows that $(I + \mathbf{G}\mathbf{K}_\varepsilon)^{-1}$ has at most finitely many poles in the compact annulus $\overline{\mathbb{E}}_\beta \setminus \mathbb{E}_1$. By (3.9), $(I + \mathbf{G}\mathbf{K}_\varepsilon)^{-1}$ does not have any poles on $\partial \mathbb{E}_1$ and so there exists $\alpha \in (\beta, 1)$ such that $(I + \mathbf{G}\mathbf{K}_\varepsilon)^{-1} \in H^\infty(\mathbb{E}_\alpha, \mathbb{C}^{p \times p})$. □

We are now in a position to prove Theorem 3.1.

Proof of Theorem 3.1. By Lemma 3.3, we know that there exists $\varepsilon^* > 0$ such that for all $\varepsilon \in (0, \varepsilon^*)$, $(I + \mathbf{G}\mathbf{K}_\varepsilon)^{-1} \in H^\infty(\mathbb{E}_1, \mathbb{C}^{p \times p})$. In the following, let $\varepsilon \in (0, \varepsilon^*)$.

We first show that the other block entries of $F(\mathbf{G}, \mathbf{K}_\varepsilon)$ are also H^∞-functions. Due to the stability of $\mathbf{G}$, it suffices to show that $\mathbf{K}_\varepsilon(I + \mathbf{G}\mathbf{K}_\varepsilon)^{-1} \in H^\infty(\mathbb{E}_1, \mathbb{C}^{m \times p})$. In the remainder of the proof, when we write $z \to \lambda_j$, it is assumed that $z \in \mathbb{E}_1$. By assumption, $\lambda_j \neq \lambda_k$ for $j, k \in \underline{N}$, $j \neq k$. Note that, by (3.2), $\mathbf{G}(\lambda_j)K_j$ is invertible. Consequently,

$$\begin{aligned}
&\lim_{z\to\lambda_j} \frac{1}{z-\lambda_j}(I + \mathbf{G}(z)\mathbf{K}_\varepsilon(z))^{-1} \\
&\qquad = \lim_{z\to\lambda_j} \left[\varepsilon\mathbf{G}(z)K_j + (z-\lambda_j)\left(I + \varepsilon\mathbf{G}(z)\mathbf{K}^0(z) + \varepsilon \sum_{k\in\underline{N},\ k\neq j} \frac{\mathbf{G}(z)K_k}{z-\lambda_k} \right) \right]^{-1} \\
&\qquad = (\varepsilon\mathbf{G}(\lambda_j)K_j)^{-1} \quad \forall j \in \underline{N}\,. \qquad (3.10)
\end{aligned}$$

By (3.1) and (3.10), we conclude that $\mathbf{K}_\varepsilon(z)(I + \mathbf{G}\mathbf{K}_\varepsilon)^{-1}$ has a finite limit at λ_j, so that $\mathbf{K}_\varepsilon(I + \mathbf{G}\mathbf{K}_\varepsilon)^{-1}$ is bounded on $\mathbb{E}_1 \cap \Lambda$, where Λ is a neighborhood of the set $\{\lambda_j : j \in \underline{N}\}$. Since $(I + \mathbf{G}\mathbf{K}_\varepsilon)^{-1} \in H^\infty(\mathbb{E}_1, \mathbb{C}^{p\times p})$ and $\mathbf{K}_\varepsilon$ is uniformly bounded on $\mathbb{E}_1 \setminus \Lambda$, it follows that $\mathbf{K}_\varepsilon(I + \mathbf{G}\mathbf{K}_\varepsilon)^{-1} \in H^\infty(\mathbb{E}_1, \mathbb{C}^{m\times p})$. Consequently, $F(\mathbf{G}, \mathbf{K}_\varepsilon) \in H^\infty(\mathbb{E}_1, \mathbb{C}^{(m+p)\times(m+p)})$, showing that $\mathcal{F}(\mathbf{G}, \mathbf{K}_\varepsilon)$ is ℓ^2-stable.

To prove that $F(\mathbf{G}, \mathbf{K}_\varepsilon) \in \hat{\ell}^1(\mathbb{C}^{(m+p)\times(m+p)})$, we set

$$\mathbf{K}^1(z) := \sum_{j=1}^{N} \frac{K_j}{z-\lambda_j}\,.$$

We see that $\mathbf{K}^1$ is a (strictly proper) rational matrix function. By a standard result (see [16, p. 75, Theorem 4.1.43]), $\mathbf{K}^1$ has a right-coprime factorization over $\mathcal{R}_{\mathrm{s}}$, i.e., $\mathbf{K}^1 = \mathbf{N}\mathbf{D}^{-1}$, where $\mathbf{N} \in \mathcal{R}_{\mathrm{s}}^{m\times p}$, $\mathbf{D} \in \mathcal{R}_{\mathrm{s}}^{p\times p}$, and there exist $\mathbf{X} \in \mathcal{R}_{\mathrm{s}}^{p\times m}$, $\mathbf{Y} \in \mathcal{R}_{\mathrm{s}}^{p\times p}$ such that $\mathbf{X}\mathbf{N} + \mathbf{Y}\mathbf{D} = I$. Therefore,

$$\mathbf{K}_\varepsilon = \varepsilon(\mathbf{K}^0 + \mathbf{K}^1) = \varepsilon(\mathbf{K}^0\mathbf{D} + \mathbf{N})\mathbf{D}^{-1}\,,$$

showing that $\mathbf{K}_\varepsilon$ has right-coprime factorization $(\varepsilon(\mathbf{K}^0\mathbf{D} + \mathbf{N}), \mathbf{D})$, since

$$(\varepsilon^{-1}\mathbf{X})\varepsilon(\mathbf{K}^0\mathbf{D} + \mathbf{N}) + (\mathbf{Y} - \mathbf{X}\mathbf{K}^0)\mathbf{D} = \mathbf{X}\mathbf{D} + \mathbf{Y}\mathbf{D} = I\,.$$

Since $\mathbf{K}^0, \mathbf{N} \in \hat{\ell}^1(\mathbb{C}^{m\times p})$ and $\mathbf{D} \in \hat{\ell}^1(\mathbb{C}^{p\times p})$, we have that $\mathbf{K}^0\mathbf{D} + \mathbf{N} \in \hat{\ell}^1(\mathbb{C}^{m\times p})$. Moreover, $(I, \mathbf{G})$ is a left-coprime factorization of $\mathbf{G}$ over $H^\infty(\mathbb{E}_1)$ and, by assumption, $\mathbf{G} \in \hat{\ell}^1(\mathbb{C}^{m\times p})$. Therefore, invoking Proposition 2.5, it follows that $F(\mathbf{G}, \mathbf{K}_\varepsilon) \in \hat{\ell}^1(\mathbb{C}^{(m+p)\times(m+p)})$.

To prove tracking and disturbance rejection, we note first that, since $\mathbf{G}(\lambda_j)K_j$ is invertible,

$$(I + \mathbf{G}\mathbf{K}_\varepsilon)^{-1}(\lambda_j) = \lim_{z\to\lambda_j} (I + \mathbf{G}(z)\mathbf{K}_\varepsilon(z))^{-1} = 0 \quad \forall j \in \underline{N} \qquad (3.11)$$

and

$$((I + \mathbf{G}\mathbf{K}_\varepsilon)^{-1}\mathbf{G})(\lambda_j) = \lim_{z\to\lambda_j} (I + \mathbf{G}(z)\mathbf{K}_\varepsilon(z))^{-1}\mathbf{G}(z) = 0 \quad \forall j \in \underline{N}\,. \qquad (3.12)$$

Let r be given by (3.4) and let d_1, d_2 satisfy (3.5). For $j \in \underline{N}$, define $a_j \in F(\mathbb{Z}_+, \mathbb{C}^p)$ and $b_j \in F(\mathbb{Z}_+, \mathbb{C}^m)$ by

$$a_j(k) := \lambda_j^k \mathfrak{r}_j\,, \quad b_j(k) := \lambda_j^k \mathfrak{d}_{1j}\,,$$

and define $\tilde{d}_1$ by

$$\tilde{d}_1(k) := d_1(k) - \sum_{j=1}^{N} b_j = d_1(k) - \sum_{j=1}^{N} \lambda_j^k \mathfrak{d}_{1j}\,.$$

Obviously, $r = \sum_{j=1}^N a_j$ and $\lim_{n\to\infty} \tilde{d}_1(k) = 0$. Then, by Lemma 2.6, (3.11), and (3.12), we obtain

$$\begin{aligned}
&\lim_{k\to\infty} [\mathscr{Z}^{-1}((I+\mathbf{G}\mathbf{K}_\varepsilon)^{-1}) \star r](k)\\
&\qquad = \sum_{j=1}^{N} \lim_{k\to\infty} \{[\mathscr{Z}^{-1}((I+\mathbf{G}\mathbf{K}_\varepsilon)^{-1}) \star a_j](k) - ((I+\mathbf{G}\mathbf{K}_\varepsilon)^{-1})(\lambda_j)\lambda_j^k \mathfrak{r}_j\}\\
&\qquad = 0
\end{aligned}\tag{3.13}$$

and

$$\begin{aligned}
&\lim_{k\to\infty} [\mathscr{Z}^{-1}((I+\mathbf{G}\mathbf{K}_\varepsilon)^{-1}\mathbf{G}) \star d_1](k)\\
&\qquad = \sum_{j=1}^{N} \lim_{k\to\infty} \{[\mathscr{Z}^{-1}((I+\mathbf{G}\mathbf{K}_\varepsilon)^{-1}\mathbf{G}) \star b_j](k) - ((I+\mathbf{G}\mathbf{K}_\varepsilon)^{-1}\mathbf{G})(\lambda_j)\lambda_j^k \mathfrak{d}_{1j}\}\\
&\qquad\qquad + \lim_{k\to\infty} [\mathscr{Z}^{-1}((I+\mathbf{G}\mathbf{K}_\varepsilon)^{-1}\mathbf{G}) \star \tilde{d}_1](k)\\
&\qquad = 0\,.
\end{aligned}\tag{3.14}$$

Similarly, by Lemma 2.6 and (3.11),

$$\lim_{k\to\infty} [\mathscr{Z}^{-1}((I+\mathbf{G}\mathbf{K}_\varepsilon)^{-1}) \star d_2](k) = 0\,.\tag{3.15}$$

By Figure 2.1 (with $\mathbf{K}$ replaced by $\mathbf{K}_\varepsilon$), it is clear that

$$\hat{r} - \hat{y} = \hat{u}_c = (I+\mathbf{G}\mathbf{K}_\varepsilon)^{-1}(\hat{r} - \hat{d}_2) - (I+\mathbf{G}\mathbf{K}_\varepsilon)^{-1}\mathbf{G}\hat{d}_1\,.\tag{3.16}$$

Therefore, by (3.13)–(3.16),

$$\begin{aligned}
\lim_{k\to\infty} (r-y)(k) = \lim_{k\to\infty} \big\{&[\mathscr{Z}^{-1}((I+\mathbf{G}\mathbf{K}_\varepsilon)^{-1}) \star (r - d_2)](k)\\
&- [\mathscr{Z}^{-1}((I+\mathbf{G}\mathbf{K}_\varepsilon)^{-1}\mathbf{G}) \star d_1](k)\big\} = 0\,.
\end{aligned}$$

This completes the proof. □

Next we show that, under a mild extra assumption on $\mathbf{G}$, $\mathbf{K}^0$, d_1, and d_2, the convergence of $y(k)$ to $r(k)$ as $k \to \infty$ is exponentially fast.

THEOREM 3.4. *Consider the discrete-time feedback system $\mathcal{F}(\mathbf{G},\mathbf{K}_\varepsilon)$ shown in Figure* 2.1 *(with $\mathbf{K}$ replaced by $\mathbf{K}_\varepsilon$). Assume that $\mathbf{G} \in H^\infty_<(\mathbb{E}_1, \mathbb{C}^{p\times m})$ and $\mathbf{K}_\varepsilon$ is given by* (3.1)*, where $\mathbf{K}^0 \in H^\infty_<(\mathbb{E}_1, \mathbb{C}^{m\times p})$, $K_j \in \mathbb{C}^{m\times p}$, and $|\lambda_j| = 1$ for $j \in \underline{N}$ with $\lambda_j \neq \lambda_k$ for $j \neq k$. If* (3.2) *holds, then there exists $\varepsilon^* > 0$ such that, for every $\varepsilon \in (0, \varepsilon^*)$, $F(\mathbf{G},\mathbf{K}_\varepsilon) \in H^\infty_<(\mathbb{E}_1, \mathbb{C}^{(m+p)\times(m+p)})$.*

Moreover, if the reference signal r is given by (3.4) *and there exist $M \geq 0$ and $\rho \in (0,1)$ such that the disturbance signals d_1, d_2 satisfy*

$$\left\| d_1(k) - \sum_{j=1}^{N} \lambda_j^k \mathfrak{d}_{1j} \right\| \leq M\rho^k\,, \quad \left\| d_2(k) - \sum_{j=1}^{N} \lambda_j^k \mathfrak{d}_{2j} \right\| \leq M\rho^k \ \ \forall k \in \mathbb{Z}_+, \tag{3.17}$$

$$\textit{where } \mathfrak{d}_{1j} \in \mathbb{C}^m,\ \mathfrak{d}_{2j} \in \mathbb{C}^p,$$

then, for every $\varepsilon \in (0, \varepsilon^*)$, there exist $L \geq 0$ and $\beta \in (\rho, 1)$ such that

$$\|y(k) - r(k)\| \leq L\beta^k \quad \forall k \in \mathbb{Z}_+\,.$$

Proof. By Lemma 3.3 and the hypotheses on $\mathbf{G}$, $\mathbf{K}^0$, we know that there exists $\varepsilon^* > 0$ such that, for every $\varepsilon \in (0, \varepsilon^*)$, there exists $\alpha \in (\rho, 1)$ such that

$$(I + \mathbf{G}\mathbf{K}_\varepsilon)^{-1} \in H^\infty(\mathbb{E}_\alpha, \mathbb{C}^{p\times p})\,,\ \mathbf{G} \in H^\infty(\mathbb{E}_\alpha, \mathbb{C}^{p\times m})\,,\ \mathbf{K}^0 \in H^\infty(\mathbb{E}_\alpha, \mathbb{C}^{m\times p})\,.$$

To prove that $F(\mathbf{G}, \mathbf{K}_\varepsilon) \in H^\infty(\mathbb{E}_\alpha, \mathbb{C}^{(m+p)\times(m+p)})$, it suffices to show that $\mathbf{K}_\varepsilon(I + \mathbf{G}\mathbf{K}_\varepsilon)^{-1} \in H^\infty(\mathbb{E}_\alpha, \mathbb{C}^{m\times p})$. By (3.10), we conclude that $\mathbf{K}_\varepsilon(z)(I + \mathbf{G}(z)\mathbf{K}_\varepsilon(z))^{-1}$ has a finite limit as $z \to \lambda_j$ for every $j \in \underline{N}$, so that $\mathbf{K}_\varepsilon(I + \mathbf{G}\mathbf{K}_\varepsilon)^{-1}$ is bounded on a neighborhood Λ of the set $\{\lambda_j : j \in \underline{N}\}$. Since $(I + \mathbf{G}\mathbf{K}_\varepsilon)^{-1} \in H^\infty(\mathbb{E}_\alpha, \mathbb{C}^{p\times p})$ and $\mathbf{K}_\varepsilon$ is uniformly bounded on $\mathbb{E}_\alpha \setminus \Lambda$, it follows that

$$\mathbf{K}_\varepsilon(I + \mathbf{G}\mathbf{K}_\varepsilon)^{-1} \in H^\infty(\mathbb{E}_\alpha, \mathbb{C}^{m\times p})\,.$$

Hence $F(\mathbf{G}, \mathbf{K}_\varepsilon) \in H^\infty(\mathbb{E}_\alpha, \mathbb{C}^{(m+p)\times(m+p)})$. Therefore, it follows from Proposition 2.7 that, for every $\beta \in (\alpha, 1)$, we have

$$(I + \mathbf{G}\mathbf{K}_\varepsilon)^{-1} \in \hat{\ell}^1_\beta(\mathbb{C}^{p\times p})\,, \quad (I + \mathbf{G}\mathbf{K}_\varepsilon)^{-1}\mathbf{G} \in \hat{\ell}^1_\beta(\mathbb{C}^{p\times m})\,.$$

Finally, invoking Lemma 2.6, (3.4), (3.11), (3.12), and (3.17), we conclude that there exists $M_1 \geq 0$ such that

$$\begin{aligned}
\|[\mathscr{Z}^{-1}((I + \mathbf{G}\mathbf{K}_\varepsilon)^{-1}) \star r](k)\| &\leq M_1\beta^k \quad \forall k \in \mathbb{Z}_+\,,\\
\|[\mathscr{Z}^{-1}((I + \mathbf{G}\mathbf{K}_\varepsilon)^{-1}\mathbf{G}) \star d_1](k)\| &\leq M_1\beta^k \quad \forall k \in \mathbb{Z}_+\,,\\
\|[\mathscr{Z}^{-1}((I + \mathbf{G}\mathbf{K}_\varepsilon)^{-1}) \star d_2](k)\| &\leq M_1\beta^k \quad \forall k \in \mathbb{Z}_+\,.
\end{aligned}$$

Consequently, by (3.16), we have

$$\|y(k) - r(k)\| \leq 3M_1\beta^k \quad \forall k \in \mathbb{Z}_+\,,$$

completing the proof. □

Application to state-space systems. We now apply Theorem 3.1 to obtain tracking results for discrete-time state-space systems. Let X be a Banach space and let the plant Σ_p be given by

$$x_p(k+1) = Ax_p(k) + Bu_p(k)\,; \quad x_p(0) = x_p^0 \in X\,, \tag{3.18a}$$

$$y_p(k) = Cx_p(k) + Du_p(k)\,, \tag{3.18b}$$

where $A \in \mathcal{B}(X, X)$, $B \in \mathcal{B}(\mathbb{C}^m, X)$, $C \in \mathcal{B}(X, \mathbb{C}^p)$, and $D \in \mathcal{B}(\mathbb{C}^m, \mathbb{C}^p)$. The transfer function $\mathbf{G}$ of Σ_p is given by

$$\mathbf{G}(z) = C(zI - A)^{-1}B + D\,.$$

Next we construct a state-space realization of the controller transfer function (3.1). Let $\mathbf{K}^0 \in \mathcal{R}_{\mathrm{s}}^{m\times p}$ and let $(A_0, B_0, C_0, D_0) \in \mathbb{C}^{n_0\times n_0} \times \mathbb{C}^{n_0\times p} \times \mathbb{C}^{m\times n_0} \times \mathbb{C}^{m\times p}$ be a stabilizable and detectable realization of $\mathbf{K}^0$; i.e., $\mathbf{K}^0(z) = C_0(zI - A_0)^{-1}B_0 + D_0$, (A_0, B_0) is stabilizable, and (C_0, A_0) is detectable. Since $\mathbf{K}^0$ is ℓ^2-stable, A_0 is power stable. Let $K_j \in \mathbb{C}^{m\times p}$ and $|\lambda_j| = 1$ for $j \in \underline{N}$ with $\lambda_j \neq \lambda_k$ for $j \neq k$. Moreover, let $A_c \in \mathbb{C}^{(Np+n_0)\times(Np+n_0)}$, $B_c \in \mathbb{C}^{(Np+n_0)\times p}$, $C_c \in \mathbb{C}^{m\times(Np+n_0)}$, and $D_c \in \mathbb{C}^{m\times p}$ be given by

$$A_c := \begin{pmatrix} A_0 & & & \\ & \lambda_1 I_p & & \\ & & \ddots & \\ & & & \lambda_N I_p \end{pmatrix}, \quad B_c := \begin{pmatrix} B_0 \\ I_p \\ \vdots \\ I_p \end{pmatrix}, \tag{3.19a}$$

$$C_c := (C_0, K_1, \dots, K_N), \quad D_c := D_0, \tag{3.19b}$$

where I_p is the $p \times p$ identity matrix. We define the controller Σ_c by

$$x_c(k+1) = A_c x_c(k) + B_c u_c(k); \quad x_c(0) = x_c^0 \in \mathbb{C}^{Np+n_0}, \tag{3.20a}$$

$$y_c(k) = \varepsilon C_c x_c(k) + \varepsilon D_c u_c(k). \tag{3.20b}$$

Obviously, the transfer function $\mathbf{K}_\varepsilon$ of Σ_c is given by

$$\mathbf{K}_\varepsilon(z) = \varepsilon(C_c(zI - A_c)^{-1}B_c + D_c) = \varepsilon\left(\mathbf{K}^0(z) + \sum_{j=1}^{N} \frac{K_j}{z - \lambda_j}\right).$$

Consider the feedback interconnection of (3.18) and (3.20) given by

$$u_c = r - y_p - d_2, \quad u_p = y_c + d_1, \quad y = y_p + d_2, \tag{3.21}$$

where r is a reference signal and d_1 and d_2 are disturbance signals. Let $\mathcal{F}(\Sigma_p, \Sigma_c)$ denote the feedback system given by (3.18)–(3.21). The state-space system $\mathcal{F}(\Sigma_p, \Sigma_c)$ is a state-space realization of the system $\mathcal{F}(\mathbf{G}, \mathbf{K}_\varepsilon)$ shown in Figure 2.1 (with $\mathbf{K}$ replaced by $\mathbf{K}_\varepsilon$).

Theorem 3.5. *Assume that* (3.18) *is power stable and that* (3.2) *holds, i.e., $\sigma(\bar{\lambda}_j \mathbf{G}(\lambda_j)K_j) \subset \mathbb{C}_0$ for every $j = \underline{N}$. Then there exists $\varepsilon^* > 0$ such that, for all $\varepsilon \in (0, \varepsilon^*)$, the following statements hold:*

1. *$\mathcal{F}(\Sigma_p, \Sigma_c)$ is power stable. Moreover, $\mathcal{F}(\Sigma_p, \Sigma_c)$ is input-to-state stable in the sense that there exist $M_1 \geq 1$ and $\gamma \in (0,1)$ such that, for all $x_p^0 \in X$, $x_c^0 \in \mathbb{C}^{Np+n_0}$, $r, d_2 \in \ell^\infty(\mathbb{Z}_+, \mathbb{C}^p)$, and all $d_1 \in \ell^\infty(\mathbb{Z}_+, \mathbb{C}^m)$,*

$$\left\|\begin{pmatrix} x_p \\ x_c \end{pmatrix}\right\|_{\ell^\infty} \leq M_1\left(\gamma^k \left\|\begin{pmatrix} x_p^0 \\ x_c^0 \end{pmatrix}\right\| + \|r\|_{\ell^\infty} + \|d_1\|_{\ell^\infty} + \|d_2\|_{\ell^\infty}\right).$$

2. *If r is given by* (3.4) *and d_1, d_2 satisfy* (3.5)*, then for all initial conditions $x_p^0 \in X$ and $x_c^0 \in \mathbb{C}^{Np+n_0}$, the output $y = y_p + d_2$ asymptotically tracks r, that is,* $\lim_{k\to\infty}(y(k) - r(k)) = 0$. *Additionally, if* (3.17) *holds with $M \geq 0$ and $\rho \in (0,1)$, then the convergence is exponentially fast.*

We omit the proof, which is based on a routine argument involving a combination of Theorem 3.1 and a result on the equivalence of input-output and power stability [5, Theorem 2]; see [4] for details.

4. Low-gain sampled-data control. In the following, let $\mathfrak{B}(\mathbb{R}_+)$ denote the Borel-σ-algebra on $\mathbb{R}_+$. For a $\mathbb{C}^{p\times m}$-valued Borel measure μ on $\mathbb{R}_+$, the total variation $|\mu| : \mathfrak{B}(\mathbb{R}_+) \to [0,\infty]$ of μ is defined by

$$|\mu|(E) := \sup\left\{ \sum_{j=1}^{\infty} \|\mu(E_j)\| : E_j \in \mathfrak{B}(\mathbb{R}_+)\,,\ E_j \cap E_k = \emptyset \text{ if } j \neq k\,,\ E = \bigcup_{j=1}^{\infty} E_j \right\}.$$

It is clear that

$$\|\mu(E)\| \le |\mu|(E) \quad \forall E \in \mathfrak{B}(\mathbb{R}_+)\,.$$

The following theorem, for which the proof is omitted, shows that a $\mathbb{C}^{p\times m}$-valued Borel measure is necessarily bounded.

THEOREM 4.1. *The total variation $|\mu|$ of a $\mathbb{C}^{p\times m}$-valued Borel measure μ is a finite nonnegative Borel measure on $\mathbb{R}_+$.*

The following technical result, for which we omit the routine proof, is used later.

PROPOSITION 4.2. *Let μ be a $\mathbb{C}^{p\times m}$-valued Borel measure on $\mathbb{R}_+$. For every $\varepsilon > 0$, there exists $T > 0$ such that*

$$\int_t^{\infty} |\mu|(ds) < \varepsilon \quad \forall t \ge T\,.$$

Let μ be a $\mathbb{C}^{p\times m}$-valued Borel measure on $\mathbb{R}_+$. Then the continuous-time input-output operator G defined by

$$(Gu)(t) := (\mu \star u)(t) = \int_0^t \mu(ds)u(t-s)\,, \quad t \ge 0\,,\ u \in L^1_{\text{loc}}(\mathbb{R}_+, \mathbb{C}^m)\,, \tag{4.1}$$

is L^q-stable for $1 \le q \le \infty$. The transfer function $\mathbf{G}$ of G is the Laplace transform of μ, that is,

$$\mathbf{G}(s) = \int_{\mathbb{R}_+} e^{-st}\mu(dt) \quad \forall s \in \overline{\mathbb{C}}_0. \tag{4.2}$$

Trivially, by Theorem 4.1, $\|\mathbf{G}(s)\| \le \int_0^\infty |\mu|(dt) < \infty$ for all $s \in \overline{\mathbb{C}}_0$. It follows that $\mathbf{G} \in H^\infty(\mathbb{C}_0, \mathbb{C}^{p\times m})$.

LEMMA 4.3. *Let the operator G be given by* (4.1)*, where μ is a $\mathbb{C}^{p\times m}$-valued Borel measure on $\mathbb{R}_+$. Then*

$$\limsup_{t\to\infty} \|(Gu)(t)\| \le |\mu|(\mathbb{R}_+) \limsup_{t\to\infty} \|u(t)\| \quad \forall u \in L_b(\mathbb{R}_+, \mathbb{C}^m)\,.$$

Proof. Let $\varepsilon > 0$. By Proposition 4.2, there exists $T > 0$ such that

$$\int_T^{\infty} |\mu|(ds) \le \frac{\varepsilon}{2\|u\|_\infty} \quad \text{and} \quad \|u(t)\| \le \sigma + \frac{\varepsilon}{2M} \quad \forall t \ge T,$$

where $M := |\mu|(\mathbb{R}_+)$ and $\sigma := \limsup_{t\to\infty} \|u(t)\|$. Hence, for $t \geq 2T$,

$$\begin{aligned}
\|(Gu)(t)\| &\leq \int_0^{t/2} \|u(t-s)\||\mu|(ds) + \int_{t/2}^{t} \|u(t-s)\||\mu|(ds) \\
&\leq \Big(\sigma + \frac{\varepsilon}{2M}\Big) \int_0^{t/2} |\mu|(ds) + \|u\|_\infty \int_{t/2}^{t} |\mu|(ds) \\
&\leq \Big(\sigma + \frac{\varepsilon}{2M}\Big) \int_0^{\infty} |\mu|(ds) + \|u\|_\infty \int_{T}^{\infty} |\mu|(ds) \\
&\leq \Big(\sigma + \frac{\varepsilon}{2M}\Big) M + \|u\|_\infty \frac{\varepsilon}{2\|u\|_\infty} \\
&\leq M\sigma + \varepsilon\,.
\end{aligned}$$

Since this holds for all $\varepsilon > 0$, the claim follows. □

LEMMA 4.4. *Let $\xi \in \overline{\mathbb{C}}_0$, $\mathfrak{v} \in \mathbb{C}^m$, $u \in L_b(\mathbb{R}_+, \mathbb{C}^m)$ and let G be given by* (4.1)*, where μ is a $\mathbb{C}^{p\times m}$-valued Borel measure on $\mathbb{R}_+$.*

1. *If $\lim_{t\to\infty}(u(t) - e^{\xi t}\mathfrak{v}) = 0$, then*

$$\lim_{t\to\infty} [(Gu)(t) - \mathbf{G}(\xi)e^{\xi t}\mathfrak{v}] = 0\,.$$

2. *If there exist $\alpha < 0$ and $M \geq 0$ such that*

$$\int_0^\infty e^{-\alpha s}|\mu|(ds) < \infty \quad \text{and} \quad \|u(t) - e^{\xi t}\mathfrak{v}\| \leq Me^{\alpha t} \quad \forall t \geq 0\,,$$

then there exists $L \geq 0$ such that

$$\|(Gu)(t) - \mathbf{G}(\xi)e^{\xi t}\mathfrak{v}\| \leq Le^{\alpha t} \quad \forall t \geq 0\,.$$

Proof. Define $v : \mathbb{R}_+ \to \mathbb{C}^m$ by $v(t) := e^{\xi t}\mathfrak{v}$. By (4.1) and (4.2), using $\xi \in \overline{\mathbb{C}}_0$, we have

$$\begin{aligned}
\|(Gu)(t) - \mathbf{G}(\xi)e^{\xi t}\mathfrak{v}\| &= \left\|\int_0^t \mu(ds)u(t-s) - \int_0^\infty e^{\xi(t-s)}\mu(ds)\mathfrak{v}\right\| \\
&\leq \left\|\int_0^t \mu(ds)(u(t-s) - e^{\xi(t-s)}\mathfrak{v})\right\| + \|\mathfrak{v}\| \int_t^\infty |e^{\xi(t-s)}||\mu|(ds) \\
&\leq \|(G(u-v))(t)\| + \|\mathfrak{v}\| \int_t^\infty |\mu|(ds) \quad \forall t \geq 0\,. \qquad (4.3)
\end{aligned}$$

By hypothesis, $\lim_{t\to\infty} \|u(t) - v(t)\| = 0$, and so, by Lemma 4.3,

$$\lim_{t\to\infty} \|G(u-v)(t)\| = 0\,. \qquad (4.4)$$

Moreover, it follows from Proposition 4.2 that $\lim_{t\to\infty} \int_t^\infty |\mu|(ds) = 0$. Hence, invoking (4.3) and (4.4) completes the proof of statement 1.

To prove statement 2, assume that there exist $\alpha < 0$ and $M \geq 0$ such that $M_1 := \int_0^\infty e^{-\alpha s}|\mu|(ds) < \infty$ and $\|u(t) - e^{\xi t}\mathfrak{v}\| \leq Me^{\alpha t}$ for all $t \geq 0$. Since $\alpha < 0$, it follows from (4.3) that

$$\begin{aligned} e^{-\alpha t}\|(Gu)(t)-\mathbf{G}(\xi)e^{\xi t}\mathfrak{v}\| &\le e^{-\alpha t}\int_0^t \|(u-v)(t-s)\||\mu|(ds)+\|\mathfrak{v}\|e^{-\alpha t}\int_t^\infty |\mu|(ds) \\ &\le M\int_0^t e^{-\alpha s}|\mu|(ds)+\|\mathfrak{v}\|\int_t^\infty e^{-\alpha s}|\mu|(ds) \\ &\le MM_1+\|\mathfrak{v}\|M_1 \quad \forall t\ge 0\,. \end{aligned}$$

Hence $\|(Gu)(t)-\mathbf{G}(\xi)e^{\xi t}\mathfrak{v}\|\le M_1(M+\|\mathfrak{v}\|)e^{\alpha t}$ for all $t\ge 0$. □

DEFINITION 4.5. *Let $\tau>0$ denote the sampling period and let $F(\mathbb{R}_+,\mathbb{C}^m)$ denote the space of all $\mathbb{C}^m$-valued functions defined on $\mathbb{R}_+$. We define the* ideal sampling operator $\mathcal{S}_\tau : F(\mathbb{R}_+,\mathbb{C}^m)\to F(\mathbb{Z}_+,\mathbb{C}^m)$ *by*

$$(\mathcal{S}_\tau u)(k) := u(k\tau) \quad \forall k\in\mathbb{Z}_+\,.$$

The (zero-order) hold operator $\mathcal{H}_\tau : F(\mathbb{Z}_+,\mathbb{C}^m)\to F(\mathbb{R}_+,\mathbb{C}^m)$ *is defined by*

$$(\mathcal{H}_\tau v)(t) := v(k) \quad \forall t\in[k\tau,(k+1)\tau)\,.$$

Define the sample-hold discretization G_τ of G by

$$G_\tau := \mathcal{S}_\tau G\mathcal{H}_\tau \tag{4.5}$$

and define $g_\tau\in F(\mathbb{Z}_+,\mathbb{C}^{p\times m})$ by

$$g_\tau(k) := \mu(E_k)\,, \quad \text{where} \quad E_k := \begin{cases} \{0\}\,, & k=0, \\ ((k-1)\tau,k\tau]\,, & k\in\mathbb{N}\,. \end{cases} \tag{4.6}$$

PROPOSITION 4.6. *Assume that G is given by (4.1) and g_τ is defined by (4.6), where μ is a $\mathbb{C}^{p\times m}$-valued Borel measure on $\mathbb{R}_+$. Then g_τ is in $\ell^1(\mathbb{Z}_+,\mathbb{C}^{p\times m})$ and the operator G_τ defined by (4.5) satisfies*

$$G_\tau v = g_\tau\star v \quad \forall v\in F(\mathbb{Z}_+,\mathbb{C}^m)\,.$$

Consequently, $G_\tau\in\mathcal{B}(\ell^q(\mathbb{Z}_+,\mathbb{C}^m),\ell^q(\mathbb{Z}_+,\mathbb{C}^p))$ for $1\le q\le\infty$.

Proof. Clearly,

$$\sum_{k=0}^\infty \|g_\tau(k)\| = \sum_{k=0}^\infty \|\mu(E_k)\| \le \sum_{k=0}^\infty |\mu|(E_k) = |\mu|(\mathbb{R}_+)<\infty\,,$$

showing that $g_\tau\in\ell^1(\mathbb{Z}_+,\mathbb{C}^{p\times m})$. For any discrete-time input $v\in F(\mathbb{Z}_+,\mathbb{C}^m)$, we have

$$\begin{aligned} (G_\tau v)(k) = ((\mathcal{S}_\tau G\mathcal{H}_\tau)v)(k) = (G(\mathcal{H}_\tau v))(k\tau) &= \int_0^{k\tau}\mu(ds)(\mathcal{H}_\tau v)(k\tau-s) \\ = \sum_{j=0}^k\int_{E_j}\mu(ds)v(k-j) &= \sum_{j=0}^k g_\tau(k)v(k-j) = (g_\tau\star v)(k) \quad \forall k\in\mathbb{Z}_+\,. \end{aligned}$$

Hence $G_\tau\in\mathcal{B}(\ell^q(\mathbb{Z}_+,\mathbb{C}^m),\ell^q(\mathbb{Z}_+,\mathbb{C}^p))$ for $1\le q\le\infty$. □

Let $\mathbf{G}_\tau$ denote the transfer function of G_τ. Note that, since $g_\tau \in \ell^1(\mathbb{Z}_+, \mathbb{C}^{p\times m})$, $\mathbf{G}_\tau(z)$ is well defined for $z \in \overline{\mathbb{E}}_1$.

Remark 4.7. Let $\alpha < 0$, assume that $\int_0^\infty e^{-\alpha t}|\mu|(dt) < \infty$, and set $\rho := e^{\alpha\tau} \in (0,1)$. Then

$$\sum_{k=0}^\infty \|g_\tau(k)\|\rho^{-k} \le e^{-\alpha\tau}\sum_{k=0}^\infty \int_{E_k} e^{-\alpha t}|\mu|(dt) = e^{-\alpha\tau}\int_0^\infty e^{-\alpha t}|\mu|(dt) < \infty\,,$$

so that $g_\tau \in \ell^1_\rho(\mathbb{Z}_+, \mathbb{C}^{p\times m})$, or, equivalently, $\mathbf{G}_\tau \in \hat{\ell}^1_\rho(\mathbb{C}^{p\times m}) \subset H^\infty(\mathbb{E}_\rho, \mathbb{C}^{p\times m})$.

LEMMA 4.8. *Let* $\xi \in \overline{\mathbb{C}}_0$. *Then* $\lim_{\tau\to 0}\mathbf{G}_\tau(e^{\xi\tau}) = \mathbf{G}(\xi)$.

Proof. Clearly,

$$\mathbf{G}_\tau(e^{\xi\tau}) = \sum_{k=0}^\infty g_\tau(k)e^{-\xi\tau k} = \sum_{k=0}^\infty \mu(E_k)e^{-\xi\tau k} = \sum_{k=0}^\infty \int_{E_k} e^{-\xi\tau k}\mu(dt)$$

and

$$\mathbf{G}(\xi) = \int_{\mathbb{R}_+} e^{-\xi t}\mu(dt) = \sum_{k=0}^\infty \int_{E_k} e^{-\xi t}\mu(dt)\,,$$

so that

$$\|\mathbf{G}_\tau(e^{\xi\tau}) - \mathbf{G}(\xi)\| = \left\|\sum_{k=0}^\infty \int_{E_k}(e^{-\xi\tau k} - e^{-\xi t})\mu(dt)\right\| \le \sum_{k=0}^\infty \int_{E_k} |e^{-\xi\tau k} - e^{-\xi t}||\mu|(dt)\,.$$

Using the fact that $\xi \in \overline{\mathbb{C}}_0$, we obtain

$$\|\mathbf{G}_\tau(e^{\xi\tau}) - \mathbf{G}(\xi)\| \le \sum_{k=0}^\infty \int_{E_k} |1 - e^{-\xi(t-\tau k)}||\mu|(dt) \le \sup_{t\in[0,\tau]} |1 - e^{\xi t}||\mu|(\mathbb{R}_+)\,.$$

Since $\lim_{\tau\to 0}\sup_{t\in[0,\tau]}|1 - e^{\xi t}| = 0$ and $|\mu|(\mathbb{R}_+)$ is finite, the claim follows. □

Remark 4.9. The convergence of $\mathbf{G}_\tau(e^{\xi\tau})$ to $\mathbf{G}(\xi)$ as $\tau \to 0$ is uniform for all $\xi \in U$ if $U \subset \overline{\mathbb{C}}_0$ is compact. Moreover, it is obvious that $\mathbf{G}_\tau(1) = \mathbf{G}(0)$ for all $\tau > 0$.

The following theorem is the main result of this section.

THEOREM 4.10. *Let* $N \in \mathbb{N}$ *and* $\xi_j \in i\mathbb{R}$ *for all* $j \in \underline{N}$ *with* $\xi_j \neq \xi_k$ *for* $j \neq k$. *Let* G *be given by* (4.1), *where* μ *is a* $\mathbb{C}^{p\times m}$*-valued Borel measure on* $\mathbb{R}_+$ *such that* $\int_0^\infty e^{-\alpha t}|\mu|(dt) < \infty$ *for some* $\alpha < 0$. *Let the discrete-time controller* $K_{\tau,\varepsilon}$ *be such that its transfer function* $\mathbf{K}_{\tau,\varepsilon}$ *is given by*

$$\mathbf{K}_{\tau,\varepsilon}(z) = \varepsilon\left(\mathbf{K}^0(z) + \sum_{j=1}^N \frac{K_j}{z - e^{\xi_j\tau}}\right)\,, \tag{4.7}$$

where $\mathbf{K}^0 \in \hat{\ell}^1(\mathbb{C}^{m\times p})$ *and* $K_j \in \mathbb{C}^{m\times p}$. *Assume that*

$$\sigma(\mathbf{G}(\xi_j)K_j) \subset \mathbb{C}_0 \quad \forall j \in \underline{N}\,. \tag{4.8}$$

The following statements hold for the output y *of the sampled-data system shown in Figure* 4.1*:*

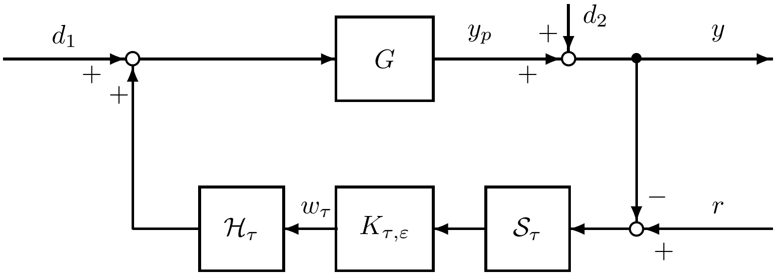


FIG. 4.1. *Sampled-data low-gain control.*

1. *There exists* $\tau^* > 0$ *such that, for every sampling period* $\tau \in (0,\tau^*)$*, there exists* $\varepsilon_\tau > 0$ *such that, for all* $\varepsilon \in (0,\varepsilon_\tau)$*, the feedback system is* L^∞*-stable, in the sense that there exists* $N_1 \geq 0$ *such that, for all* $r, d_2 \in L_b(\mathbb{R}_+,\mathbb{C}^p)$ *and all* $d_1 \in L_b(\mathbb{R}_+,\mathbb{C}^m)$*,*

$$\|y\|_\infty \leq N_1(\|r\|_\infty + \|d_1\|_\infty + \|d_2\|_\infty)\,.$$

If r *is given by*

$$r(t) := \sum_{j=1}^{N} e^{\xi_j t}\mathfrak{r}_j \quad \forall t \geq 0 \;\; \textit{where } \mathfrak{r}_j \in \mathbb{C}^p\,, \tag{4.9}$$

and $d_1 \in L_b(\mathbb{R}_+,\mathbb{C}^m)$*,* $d_2 \in L_b(\mathbb{R}_+,\mathbb{C}^p)$ *satisfy*

$$\lim_{t\to\infty}\left(d_1(t) - \sum_{j=1}^{N} e^{\xi_j t}\mathfrak{d}_{1j}\right) = 0\,,\; \lim_{t\to\infty}\left(d_2(t) - \sum_{j=1}^{N} e^{\xi_j t}\mathfrak{d}_{2j}\right) = 0\,, \tag{4.10}$$

where $\mathfrak{d}_{1j} \in \mathbb{C}^m$*,* $\mathfrak{d}_{2j} \in \mathbb{C}^p$*,*

then, for every $\delta > 0$*, there exists* $\tau_\delta \in (0,\tau^*)$ *such that, for every sampling period* $\tau \in (0,\tau_\delta)$ *and every* $\varepsilon \in (0,\varepsilon_\tau)$*,*

$$\limsup_{t\to\infty}\|y(t) - r(t)\| \leq \delta\,. \tag{4.11}$$

2. *Under the additional assumptions that* $\mathbf{K}^0 \in H^\infty_<(\mathbb{E}_1,\mathbb{C}^{m\times p})$ *and that there exist* $\gamma \in (\alpha,0)$ *and* $N_2 \geq 0$ *such that*

(4.12)
$$\left\|d_1(t) - \sum_{j=1}^{N} e^{\xi_j t}\mathfrak{d}_{1j}\right\| \leq N_2 e^{\gamma t}\,, \quad \left\|d_2(t) - \sum_{j=1}^{N} e^{\xi_j t}\mathfrak{d}_{2j}\right\| \leq N_2 e^{\gamma t} \quad \forall t \geq 0\,,$$

(4.11) *can be replaced by*

$$\|y(t) - r(t)\| \leq \delta + N_3 e^{\beta t} \quad \forall t \geq 0$$

for suitable $\beta \in (\gamma,0)$ *and* $N_3 \geq 0$ *(both depending on* τ *and* ε*).*

Only very little plant information is required in order to apply Theorem 4.10, namely, stability of the system to be controlled and some information on $\mathbf{G}(\xi_j)$, where the latter is required for the computation of K_j such that (4.8) holds. The spectral condition (4.8) is robust with respect to "sufficiently small" plant perturbations.

Proof of Theorem 4.10. To prove statement 1, set $\tau_0 := 2\pi/\sup\{|\xi_j - \xi_k| : j,k \in \underline{N},\, j \neq k\}$ and note that if $\tau \in (0,\tau_0)$, then $e^{\xi_j\tau} \neq e^{\xi_k\tau}$ for all $j,k \in \underline{N}$, $j \neq k$. It follows from Lemma 4.8 that

$$\lim_{\tau\to 0} e^{\bar{\xi}_j\tau}\mathbf{G}_\tau(e^{\xi_j\tau})K_j = \mathbf{G}(\xi_j)K_j \quad \forall j \in \underline{N}\,.$$

Hence, by hypothesis (4.8), there exists $\tau^* \in (0,\tau_0)$ such that

$$\sigma(e^{\bar{\xi}_j\tau}\mathbf{G}_\tau(e^{\xi_j\tau})K_j) \subset \mathbb{C}_0 \quad \forall j \in \underline{N}\,,\ \forall \tau \in (0,\tau^*)\,. \tag{4.13}$$

By assumption, there exists $\alpha < 0$ such that $\int_0^\infty e^{-\alpha t}|\mu|(dt) < \infty$. Therefore, by Remark 4.7, $\mathbf{G}_\tau \in H^\infty_<(\mathbb{E}_1,\mathbb{C}^{p\times m})$. Clearly,

$$\limsup_{z\to e^{\xi_j\tau},\, z\in\mathbb{E}_1} \left\| \frac{\mathbf{G}_\tau(z) - \mathbf{G}_\tau(e^{\xi_j\tau})}{z - e^{\xi_j\tau}} \right\| < \infty$$

holds for every $j \in \underline{N}$. Moreover, by assumption, $\mathbf{K}^0 \in \hat{\ell}^1(\mathbb{C}^{m\times p})$. It follows from Theorem 3.1 that, for every $\tau \in (0,\tau^*)$, there exists $\varepsilon_\tau > 0$ such that, for all $\varepsilon \in (0,\varepsilon_\tau)$,

$$\mathbf{K}_{\tau,\varepsilon}(I + \mathbf{G}_\tau\mathbf{K}_{\tau,\varepsilon})^{-1} \in \hat{\ell}^1(\mathbb{C}^{m\times p})\,.$$

Consequently, for all such τ and ε, the convolution operator $K_{\tau,\varepsilon}(I + G_\tau K_{\tau,\varepsilon})^{-1}$ has impulse response in $\ell^1(\mathbb{Z}_+,\mathbb{C}^{m\times p})$.

In the following, let $\tau \in (0,\tau^*)$ and $\varepsilon \in (0,\varepsilon_\tau)$. Set

$$M := |\mu|(\mathbb{R}_+) \qquad \text{and} \qquad M_1 := \|K_{\tau,\varepsilon}(I + G_\tau K_{\tau,\varepsilon})^{-1}\|\,. \tag{4.14}$$

Let $d_1 \in L_b(\mathbb{R}_+,\mathbb{C}^m)$ and $d_2, r \in L_b(\mathbb{R}_+,\mathbb{C}^p)$. It is well known that $\|Gd_1\|_\infty \leq M\|d_1\|_\infty$. Furthermore, set

$$d := Gd_1 + d_2\,. \tag{4.15}$$

Trivially,

$$\|\mathcal{S}_\tau d\|_{\ell^\infty} \leq \|d\|_\infty \leq M\|d_1\|_\infty + \|d_2\|_\infty \qquad \text{and} \qquad \|\mathcal{S}_\tau r\|_{\ell^\infty} \leq \|r\|_\infty\,. \tag{4.16}$$

The discrete-time signal w_τ in Figure 4.1 is given by

$$w_\tau = K_{\tau,\varepsilon}\mathcal{S}_\tau[r - (G\mathcal{H}_\tau w_\tau + d)] = K_{\tau,\varepsilon}[\mathcal{S}_\tau r - (G_\tau w_\tau + \mathcal{S}_\tau d)]\,.$$

It follows that

$$w_\tau = K_{\tau,\varepsilon}(I + G_\tau K_{\tau,\varepsilon})^{-1}(\mathcal{S}_\tau r - \mathcal{S}_\tau d)\,. \tag{4.17}$$

Invoking (4.14) and (4.16), we have

$$\|w_\tau\|_{\ell^\infty} \leq M_1(\|\mathcal{S}_\tau r\|_{\ell^\infty} + \|\mathcal{S}_\tau d\|_{\ell^\infty}) \leq M_1(\|r\|_\infty + M\|d_1\|_\infty + \|d_2\|_\infty)\,. \tag{4.18}$$

Clearly, the continuous-time signal y in Figure 4.1 satisfies

$$y = G\mathcal{H}_\tau w_\tau + Gd_1 + d_2 = G\mathcal{H}_\tau w_\tau + d\,. \tag{4.19}$$

Since $\|\mathcal{H}_\tau w_\tau\|_\infty = \|w_\tau\|_{\ell^\infty}$, it follows from (4.18) and (4.19) that

$$\begin{aligned}
\|y\|_\infty &\le \|G\mathcal{H}_\tau w_\tau\|_\infty + \|Gd_1\|_\infty + \|d_2\|_\infty \\
&\le M\|\mathcal{H}_\tau w_\tau\|_\infty + M\|d_1\|_\infty + \|d_2\|_\infty \\
&= M\|w_\tau\|_{\ell^\infty} + M\|d_1\|_\infty + \|d_2\|_\infty \\
&\le MM_1(\|r\|_\infty + M\|d_1\|_\infty + \|d_2\|_\infty) + M\|d_1\|_\infty + \|d_2\|_\infty \\
&\le N_1(\|r\|_\infty + \|d_1\|_\infty + \|d_2\|_\infty)\,,
\end{aligned}$$

with $N_1 := (M+1)(MM_1+1)$. This completes the proof of the L^∞-stability of the feedback system.

To prove approximate tracking (see (4.11)), note that, by (4.13), $\mathbf{G}_\tau(e^{\xi_j\tau})K_j$ is invertible for every $j \in \underline{N}$ and every $\tau \in (0,\tau^*)$. In the following, we take limits as $z \to e^{\xi_j\tau}$ for $z \in \mathbb{E}_1$. It is assumed that $\tau \in (0,\tau^*)$ and $\varepsilon \in (0,\varepsilon_\tau)$. A straightforward calculation yields that

$$\lim_{z\to e^{\xi_j\tau}} (I + \mathbf{G}_\tau(z)\mathbf{K}_{\tau,\varepsilon}(z))^{-1} = 0 \quad \forall j \in \underline{N}$$

and

$$\lim_{z\to e^{\xi_j\tau}} \frac{1}{z - e^{\xi_j\tau}} (I + \mathbf{G}_\tau(z)\mathbf{K}_{\tau,\varepsilon}(z))^{-1} = (\varepsilon\mathbf{G}_\tau(e^{\xi_j\tau})K_j)^{-1} \quad \forall j \in \underline{N}\,.$$

Consequently,

$$\begin{aligned}
(\mathbf{K}_{\tau,\varepsilon}(I + \mathbf{G}_\tau\mathbf{K}_{\tau,\varepsilon})^{-1})(e^{\xi_j\tau}) &= \lim_{z\to e^{\xi_j\tau}} \varepsilon\mathbf{K}^0(z)(I + \mathbf{G}_\tau(z)\mathbf{K}_{\tau,\varepsilon}(z))^{-1} \\
&\quad + \lim_{z\to e^{\xi_j\tau}} \sum_{k=1}^{N} \left(\frac{\varepsilon K_k}{z - e^{\xi_k\tau}} (I + \mathbf{G}_\tau(z)\mathbf{K}_{\tau,\varepsilon}(z))^{-1} \right) \\
&= \lim_{z\to e^{\xi_j\tau}} \frac{\varepsilon K_j}{z - e^{\xi_j\tau}} (I + \mathbf{G}_\tau(z)\mathbf{K}_{\tau,\varepsilon}(z))^{-1} \\
&= K_j(\mathbf{G}_\tau(e^{\xi_j\tau})K_j)^{-1} \quad \forall j \in \underline{N}\,. \qquad (4.20)
\end{aligned}$$

Setting

$$\mathfrak{d}_j := \mathbf{G}(\xi_j)\mathfrak{d}_{1j} + \mathfrak{d}_{2j} \quad \forall j \in \underline{N}\,, \tag{4.21}$$

it follows from the definition of d (see (4.15)) that

$$d(t) - \sum_{j=1}^{N} e^{\xi_j t}\mathfrak{d}_j = (Gd_1)(t) - \sum_{j=1}^{N} e^{\xi_j t}\mathbf{G}(\xi_j)\mathfrak{d}_{1j} + d_2(t) - \sum_{j=1}^{N} e^{\xi_j t}\mathfrak{d}_{2j}\,.$$

Invoking Lemma 4.4 and (4.10), we obtain that

$$\lim_{t\to\infty} \left\| d(t) - \sum_{j=1}^{N} e^{\xi_j t}\mathfrak{d}_j \right\| = 0. \tag{4.22}$$

It follows trivially from (4.9) and (4.22) that

$$(\mathcal{S}_\tau r)(k) = \sum_{j=1}^N e^{\xi_j k\tau}\mathfrak{r}_j \ \forall k \in \mathbb{Z}_+ \ \text{ and } \ \lim_{k\to\infty}\left\|(\mathcal{S}_\tau d)(k) - \sum_{j=1}^N e^{\xi_j k\tau}\mathfrak{d}_j\right\| = 0\,. \tag{4.23}$$

Define $a_\tau, b_\tau \in F(\mathbb{Z}_+, \mathbb{C}^m)$ by

$$a_\tau(k) := \sum_{j=1}^N e^{\xi_j\tau k}K_j(\mathbf{G}_\tau(e^{\xi_j\tau})K_j)^{-1}\mathfrak{r}_j\,, \quad b_\tau(k) := \sum_{j=1}^N e^{\xi_j\tau k}K_j(\mathbf{G}_\tau(e^{\xi_j\tau})K_j)^{-1}\mathfrak{d}_j\,. \tag{4.24}$$

It follows from Lemma 2.6, (4.17), (4.20), and (4.23) that

$$\lim_{k\to\infty}[w_\tau(k) - a_\tau(k) + b_\tau(k)] = 0\,. \tag{4.25}$$

By (4.8), $\mathbf{G}(\xi_j)K_j$ is invertible for every $j \in \underline{N}$. Define $v_1, v_2 : \mathbb{R}_+ \to \mathbb{C}^m$ by

$$v_1(t) := \sum_{j=1}^N e^{\xi_j t}K_j(\mathbf{G}(\xi_j)K_j)^{-1}\mathfrak{r}_j\,, \quad v_2(t) := \sum_{j=1}^N e^{\xi_j t}K_j(\mathbf{G}(\xi_j)K_j)^{-1}\mathfrak{d}_j\,. \tag{4.26}$$

We conclude from Lemma 4.4 and (4.9) that

$$\lim_{t\to\infty}[(Gv_1)(t) - r(t)] = \sum_{j=1}^N \lim_{t\to\infty}[(G(e^{\xi_j\cdot}K_j(\mathbf{G}(\xi_j)K_j)^{-1}\mathfrak{r}_j))(t) - e^{\xi_j t}\mathfrak{r}_j] = 0\,. \tag{4.27}$$

Furthermore, writing

$$(Gv_2)(t) - d(t) = \sum_{j=1}^N \left[(G(e^{\xi_j\cdot}K_j(\mathbf{G}(\xi_j)K_j)^{-1}\mathfrak{d}_j))(t) - e^{\xi_j t}\mathfrak{d}_j\right] + \sum_{j=1}^N e^{\xi_j t}\mathfrak{d}_j - d(t)\,,$$

an application of Lemma 4.4, and (4.22) yields that

$$\lim_{t\to\infty}[(Gv_2)(t) - d(t)] = 0. \tag{4.28}$$

Let $\delta > 0$. Invoking Lemma 4.8 and the fact that $\xi_j \in i\mathbb{R}$, there exists $\tau_\delta \in (0, \tau^*)$ such that if $\tau \in (0, \tau_\delta)$, then

$$\begin{aligned}
&\sup_{t\in[k\tau,(k+1)\tau)} \|v_1(t) - (\mathcal{H}_\tau a_\tau)(t)\| \\
&= \sup_{t\in[k\tau,(k+1)\tau)} \left\|\sum_{j=1}^N e^{\xi_j t}K_j(\mathbf{G}(\xi_j)K_j)^{-1}\mathfrak{r}_j - \sum_{j=1}^N e^{\xi_j\tau k}K_j(\mathbf{G}_\tau(e^{\xi_j\tau})K_j)^{-1}\mathfrak{r}_j\right\| \\
&\le \sup_{t\in[k\tau,(k+1)\tau)} \sum_{j=1}^N |e^{\xi_j(t-k\tau)} - 1|\|K_j(\mathbf{G}(\xi_j)K_j)^{-1}\mathfrak{r}_j\| \\
&\qquad + \sup_{t\in[k\tau,(k+1)\tau)} \sum_{j=1}^N (\|K_j\|\|(\mathbf{G}(\xi_j)K_j)^{-1} - (\mathbf{G}_\tau(e^{\xi_j\tau})K_j)^{-1}\|\|\mathfrak{r}_j\|) \\
&\le \frac{\delta}{2M} \quad \forall k \in \mathbb{Z}_+\,,
\end{aligned}$$

and, similarly,

$$\sup_{t\in[k\tau,(k+1)\tau)} \|v_2(t) - (\mathcal{H}_\tau b_\tau)(t)\| \le \frac{\delta}{2M} \quad \forall k \in \mathbb{Z}_+\,,$$

where M is defined in (4.14). Hence,

$$\sup_{t\ge 0} \|v_1(t) - (\mathcal{H}_\tau a_\tau)(t)\| + \sup_{t\ge 0} \|v_2(t) - (\mathcal{H}_\tau b_\tau)(t)\| \le \frac{\delta}{M}\,. \tag{4.29}$$

Let $\tau \in (0, \tau_\delta)$ and $\varepsilon \in (0, \varepsilon_\tau)$. Then, writing

$$\mathcal{H}_\tau w_\tau - v_1 + v_2 = \mathcal{H}_\tau(w_\tau - a_\tau + b_\tau) + (\mathcal{H}_\tau a_\tau - v_1) + (v_2 - \mathcal{H}_\tau b_\tau)$$

and invoking (4.25) and (4.29), we obtain

$$\limsup_{t\to\infty} \|(\mathcal{H}_\tau w_\tau)(t) - v_1(t) + v_2(t)\| \le \frac{\delta}{M}\,. \tag{4.30}$$

By (4.19),

$$y - r = G(\mathcal{H}_\tau w_\tau - v_1 + v_2) + (d - Gv_2) + (Gv_1 - r)\,,$$

so that it follows from (4.27) and (4.28) that

$$\limsup_{t\to\infty} \|y(t) - r(t)\| \le \limsup_{t\to\infty} \|(G(\mathcal{H}_\tau w_\tau - v_1 + v_2))(t)\|\,.$$

Finally, $\mathcal{H}_\tau w_\tau - v_1 + v_2$ is bounded and thus, by Lemma 4.3 and (4.30),

$$\limsup_{t\to\infty} \|y(t) - r(t)\| \le M \limsup_{t\to\infty} \|(\mathcal{H}_\tau w_\tau)(t) - v_1(t) + v_2(t)\| \le \delta\,,$$

completing the proof of statement 1.

To prove statement 2 of Theorem 4.10, let $\tau \in (0, \tau_\delta)$ and $\varepsilon \in (0, \varepsilon_\tau)$. Assume that $\mathbf{K}^0 \in H^\infty_<(\mathbb{E}_1, \mathbb{C}^{m\times p})$ and that there exist $N_2 \ge 0$ and $\gamma \in (\alpha, 0)$ such that (4.12) holds. Invoking Remark 4.7, we conclude that $\mathbf{G}_\tau \in H^\infty_<(\mathbb{E}_1, \mathbb{C}^{p\times m})$. Therefore, by Theorem 3.4, $\mathbf{K}_{\tau,\varepsilon}(I + \mathbf{G}_\tau \mathbf{K}_{\tau,\varepsilon})^{-1} \in H^\infty_<(\mathbb{E}_1, \mathbb{C}^{m\times p})$. Hence, by Proposition 2.7, there exists $\rho \in (e^{\gamma\tau}, 1)$ such that

$$\mathbf{K}_{\tau,\varepsilon}(I + \mathbf{G}_\tau \mathbf{K}_{\tau,\varepsilon})^{-1} \in \hat{\ell}^1_\rho(\mathbb{C}^{m\times p})\,.$$

By Lemma 4.4 and (4.12), there exists $M_2 \ge 0$ such that

$$\left\|(Gd_1)(t) - \sum_{j=1}^N e^{\xi_j t}\mathbf{G}(\xi_j)\mathfrak{d}_{1j}\right\| \le M_2 e^{\gamma t} \quad \forall t \ge 0\,.$$

Invoking (4.12), it follows that

$$\begin{aligned}\left\|d(t) - \sum_{j=1}^N e^{\xi_j t}\mathfrak{d}_j\right\| &\le \left\|(Gd_1)(t) - \sum_{j=1}^N e^{\xi_j t}\mathbf{G}(\xi_j)\mathfrak{d}_{1j}\right\| + \left\|d_2(t) - \sum_{j=1}^N e^{\xi_j t}\mathfrak{d}_{2j}\right\| \\ &\le (M_2 + N_2)e^{\gamma t} \quad \forall t \ge 0\,, \end{aligned} \tag{4.31}$$

where d and $\mathfrak{d}_j$ are defined in (4.15) and (4.21), respectively. Trivially,

$$\left\|(\mathcal{S}_\tau d)(k) - \sum_{j=1}^{N} e^{\xi_j k\tau}\mathfrak{d}_j\right\| \le (M_2+N_2)(e^{\gamma\tau})^k \le (M_2+N_2)\rho^k \quad \forall k \in \mathbb{Z}_+ .$$

It follows from (4.20) and Lemma 2.6 that there exists $M_3 \ge 0$ such that

$$\|w_\tau(k) - a_\tau(k) + b_\tau(k)\| \le M_3\rho^k \quad \forall k \in \mathbb{Z}_+ , \tag{4.32}$$

where w_τ and a_τ, b_τ are defined in (4.17) and (4.24), respectively. We conclude from Lemma 4.4, (4.9), and (4.31) that there exists $M_4 \ge 0$ such that

$$\|(Gv_1)(t) - r(t)\| \le M_4 e^{\gamma t}, \quad \|(Gv_2)(t) - d(t)\| \le M_4 e^{\gamma t}; \quad \forall t \ge 0 , \tag{4.33}$$

where v_1 and v_2 are defined in (4.26). Since $\rho \in (0,1)$, we have

$$\rho^k \le \rho^{-1}\rho^{(k\tau+\theta)/\tau} = \rho^{-1}e^{\beta(k\tau+\theta)} \quad \forall \theta \in [0,\tau) , \ \forall k \in \mathbb{Z}_+ ,$$

where $\beta := (\ln\rho)/\tau$. Consequently, by (4.32) and (4.29),

$$\begin{aligned} \|(\mathcal{H}_\tau w_\tau)(t) - v_1(t) + v_2(t)\| &\le \|(\mathcal{H}_\tau w_\tau - \mathcal{H}_\tau a_\tau + \mathcal{H}_\tau b_\tau)(t)\| \\ &\quad + \|(\mathcal{H}_\tau a_\tau)(t) - v_1(t)\| + \|v_2(t) - (\mathcal{H}_\tau b_\tau)(t)\| \\ &\le M_3\rho^{-1}e^{\beta t} + \frac{\delta}{M} \quad \forall t \ge 0 . \end{aligned}$$

Since $\rho \in (e^{\gamma\tau}, 1)$, we have that $\beta \in (\gamma, 0) \subset (\alpha, 0)$, and hence

$$\begin{aligned} \|(G(\mathcal{H}_\tau w_\tau - v_1 + v_2))(t)\| &\le \int_0^t \|(\mathcal{H}_\tau w_\tau - v_1 + v_2)(t-s)\||\mu|(ds) \\ &\le \int_0^t M_3\rho^{-1}e^{\beta(t-s)}|\mu|(ds) + \frac{\delta}{M}\int_0^\infty |\mu|(ds) \\ &\le M_3\rho^{-1}e^{\beta t}\int_0^\infty e^{-\beta s}|\mu|(ds) + \delta \\ &\le M_3M_5\rho^{-1}e^{\beta t} + \delta \quad \forall t \ge 0 , \end{aligned}$$

where $M_5 := \int_0^\infty e^{-\beta s}|\mu|(ds) \le \int_0^\infty e^{-\alpha s}|\mu|(ds) < \infty$. Therefore, by (4.19) and (4.33), it follows that

$$\begin{aligned} \|y(t) - r(t)\| &\le \|(G(\mathcal{H}_\tau w_\tau - v_1 + v_2))(t)\| + \|(d(t) - Gv_2(t))\| + \|(Gv_1)(t) - r(t)\| \\ &\le \|(G(\mathcal{H}_\tau w_\tau - v_1 + v_2))(t)\| + 2M_4e^{\gamma t} \\ &\le \delta + (M_3M_5\rho^{-1} + 2M_4)e^{\beta t} \quad \forall t \ge 0 . \end{aligned}$$

This completes the proof. □

Remark 4.11. The proof of Theorem 4.10 shows that, for fixed $\{\xi_j : j \in \underline{N}\}$, τ_δ and ε_τ can be chosen to be uniform for all signals r, d_1, and d_2 with $\mathfrak{r}_j$, $\mathfrak{d}_{1j}$, and $\mathfrak{d}_{2j}$, $j \in \underline{N}$, satisfying a prespecified bound.

Application to state-space systems. In the following, we apply the input-output results in this paper to a class of infinite-dimensional state-space systems.

Let X be a Hilbert space and assume that the plant is given by

$$\dot{x}_p(t) = Ax_p(t) + Bu_p(t)\,; \quad x_p(0) = x_p^0 \in X\,, \tag{4.34a}$$

$$y_p(t) = Cx_p(t) + Du_p(t)\,, \tag{4.34b}$$

where $A : D(A) \to X$ is the generator of a strongly continuous semigroup $\mathbf{T}(t)$ on X, $B \in \mathcal{B}(\mathbb{C}^m, X_{-1})$ is the control operator, $C \in \mathcal{B}(X, \mathbb{C}^p)$ is the (bounded) observation operator, and $D \in \mathbb{C}^{p\times m}$ is the feedthrough matrix. Here X_{-1} is the completion of X with respect to the norm $\|x\|_{-1} := \|(\beta I - A)^{-1}x\|_X$, where β is in the resolvent set A. It is known that X_{-1} does not depend on the choice of β. Moreover, $X \hookrightarrow X_{-1}$ and $\mathbf{T}(t)$ extends to a C_0-semigroup on X_{-1}. The generator of the extended semigroup is a bounded operator from X to X_{-1} which extends A. The extended semigroup and its generator will be denoted by the same symbols $\mathbf{T}(t)$ and A, respectively. We assume that B is admissible for $\mathbf{T}(t)$, that is, for every $t \geq 0$, there exists $b_t \geq$ such that

$$\left\| \int_0^t \mathbf{T}(t-s)Bu(s) \right\|_X \leq b_t \|u\|_{L^2} \quad \forall u \in L^2([0,t], \mathbb{C}^m)\,.$$

The admissibility assumption implies, in particular, that system (4.34) is regular (see [15, 18] for more details on admissible control operators and regular systems). For $u_p \in L^2_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{C}^m)$, the mild solution x_p of (4.34a), given by

$$x_p(t) = \mathbf{T}(t)x_p^0 + \int_0^t \mathbf{T}(t-\sigma)Bu_p(\sigma)d\sigma\,, \tag{4.35}$$

is a continuous X-valued function, satisfying the differential equation (4.34a) in X_{-1} for almost every $t \in \mathbb{R}_+$. The transfer function $\mathbf{G}$ of (4.34) is given by

$$\mathbf{G}(s) = C(sI - A)^{-1}B + D \quad \forall s \in \mathbb{C}_{\omega(\mathbf{T})}\,,$$

where

$$\omega(\mathbf{T}) := \lim_{t\to\infty} \frac{1}{t} \ln \|\mathbf{T}(t)\|\,.$$

We say that (4.34) is *exponentially stable* if $\omega(\mathbf{T}) < 0$. Let $\mathbf{K}^0 \in \mathcal{R}_{\mathrm{s}}^{m\times p}$ and let $(A_0, B_0, C_0, D_0) \in \mathbb{C}^{n_0\times n_0} \times \mathbb{C}^{n_0\times p} \times \mathbb{C}^{m\times n_0} \times \mathbb{C}^{m\times p}$ be a stabilizable and detectable realization of $\mathbf{K}^0$; i.e., $\mathbf{K}^0(z) = C_0(zI - A_0)^{-1}B_0 + D_0$, (A_0, B_0) is stabilizable, and (C_0, A_0) is detectable. Since $\mathbf{K}^0$ is ℓ^2-stable, it follows that A_0 is power stable. Let $A_c \in \mathbb{C}^{(Np+n_0)\times(Np+n_0)}$, $B_c \in \mathbb{C}^{(Np+n_0)\times p}$, $C_c \in \mathbb{C}^{m\times(Np+n_0)}$, and $D_c \in \mathbb{C}^{m\times p}$ be given by (3.19) with $\lambda_j = e^{\xi_j\tau}$, $\xi_j \in i\mathbb{R}$ for $j \in \underline{N}$. We define the controller by

$$x_c(k+1) = A_c x_c(k) + B_c u_c(k)\,; \quad x_c(0) = x_c^0 \in \mathbb{C}^{Np+n_0}\,, \tag{4.36a}$$

$$y_c(k) = \varepsilon C_c x_c(k) + \varepsilon D_c u_c(k)\,. \tag{4.36b}$$

The transfer function $\mathbf{K}_{\tau,\varepsilon}$ of (4.36) is given by

$$\mathbf{K}_{\tau,\varepsilon}(z) = \varepsilon \left(\mathbf{K}^0(z) + \sum_{j=1}^{N} \frac{K_j}{z - e^{\xi_j\tau}} \right)\,.$$

We consider the following feedback interconnection of (4.34) and (4.36):

$$u_p = \mathcal{H}_\tau y_c + d_1\,, \quad y = y_p + d_2\,, \quad u_c = \mathcal{S}_\tau(r - y)\,, \tag{4.37}$$

where r is a reference signal and d_1 and d_2 are disturbance signals.

THEOREM 4.12. *Consider the sampled-data state-space system given by* (4.34), (4.36), *and* (4.37). *Assume that* (4.34) *is exponentially stable and* $\sigma(\mathbf{G}(\xi_j)K_j) \subset \mathbb{C}_0$ *for all* $j = \underline{N}$. *The following statements hold:*

1. *There exists* $\tau^* > 0$ *such that, for every sampling period* $\tau \in (0, \tau^*)$, *there exists* $\varepsilon_\tau > 0$ *such that if* $\varepsilon \in (0, \varepsilon_\tau)$, *then the sampled-data system is exponentially stable; i.e., for every* $\varepsilon \in (0, \varepsilon_\tau)$, *there exist* $N_1 \geq 0$ *and* $\beta < 0$ *such that*

$$\left\| \begin{pmatrix} x_p(k\tau + \theta) \\ x_c(k) \end{pmatrix} \right\| \leq N_1 \left(e^{\beta(k\tau+\theta)} \left\| \begin{pmatrix} x_p^0 \\ x_c^0 \end{pmatrix} \right\| + \|r\|_\infty + \|d_1\|_\infty + \|d_2\|_\infty \right)$$
$$\forall \theta \in [0, \tau)\,,\ \forall k \in \mathbb{Z}_+\,,\ \forall x_p^0 \in X\,,\ \forall x_c^0 \in \mathbb{C}^{Np+n_0}\,,$$
$$\forall r, d_2 \in L_b(\mathbb{R}_+, \mathbb{C}^p)\,,\ \forall d_1 \in L_b(\mathbb{R}_+, \mathbb{C}^m)\,.$$

2. *If* r *is of the form* (4.9) *and* $d_1 \in L_b(\mathbb{R}_+, \mathbb{C}^m)$, $d_2 \in L_b(\mathbb{R}_+, \mathbb{C}^p)$ *satisfy* (4.10), *then, for every* $\delta > 0$, *there exists* $\tau_\delta > 0$ *such that, for every sampling period* $\tau \in (0, \tau_\delta)$, *there exists* $\varepsilon_\tau > 0$, *such that, for every* $\varepsilon \in (0, \varepsilon_\tau)$,

$$\limsup_{t\to\infty} \|y(t) - r(t)\| \leq \delta \quad \forall x_p^0 \in X\,,\ x_c^0 \in \mathbb{C}^{Np+n_0}\,.$$

Proof. The sample-hold discretization of (4.34) is given by the quadruple

$$\left(\mathbf{T}(\tau)\,,\ \int_0^\tau \mathbf{T}(s)Bds\,,\ C\,,\ D \right)\,. \tag{4.38}$$

Clearly, since $\mathbf{T}(t)$ is exponentially stable, $\mathbf{T}(\tau)$ is power stable. Since admissibility of B for $\mathbf{T}(t)$ implies that $A^{-1}B \in \mathcal{B}(\mathbb{C}^m, X)$ and

$$\int_0^\tau \mathbf{T}(s)Bvds = (\mathbf{T}(\tau) - I)A^{-1}Bv \quad \forall v \in \mathbb{C}^m\,,$$

we see that $\int_0^\tau \mathbf{T}(s)Bds \in \mathcal{B}(\mathbb{C}^m, X)$ for every $\tau > 0$. The transfer function of (4.38) is denoted by $\mathbf{G}_\tau$. By Lemma 4.8 and the assumption that $\sigma(\mathbf{G}(\xi_j)K_j) \subset \mathbb{C}_0$, there exists $\tau^* > 0$ such that if $\tau \in (0, \tau^*)$, then $e^{\xi_j\tau} \neq e^{\xi_k\tau}$ for all $j, k \in \underline{N}$, $j \neq k$, and

$$\sigma(e^{\bar{\xi}_j\tau}\mathbf{G}_\tau(e^{\xi_j\tau})K_j) \subset \mathbb{C}_0 \quad \forall j \in \underline{N}\,. \tag{4.39}$$

Define

$$E := (I + \varepsilon D_c D)^{-1}\,, \quad E_c := (I + \varepsilon D D_c)^{-1}\,,$$

and $\Delta : [0, \tau] \to \mathcal{B}(X \times \mathbb{C}^{Np+n_0})$ by

$$\Delta(\theta) := \begin{pmatrix} \mathbf{T}(\theta) & 0 \\ 0 & A_c \end{pmatrix} + \begin{pmatrix} \int_0^\theta \mathbf{T}(s)Bds & 0 \\ 0 & B_c \end{pmatrix} \begin{pmatrix} E & 0 \\ 0 & E_c \end{pmatrix} \begin{pmatrix} -\varepsilon D_c & \varepsilon I \\ -I & -\varepsilon D \end{pmatrix} \begin{pmatrix} C & 0 \\ 0 & C_c \end{pmatrix}.$$

For $\theta \in [0,\tau]$ and $k \in \mathbb{Z}_+$, define $R(k,\theta) : L_b(\mathbb{R}_+,\mathbb{C}^m) \times L_b(\mathbb{R}_+,\mathbb{C}^p) \times L_b(\mathbb{R}_+,\mathbb{C}^p) \to X \times \mathbb{C}^{Np+n_0}$ by

$$R(k,\theta)\begin{pmatrix} d_1 \\ d_2 \\ r \end{pmatrix} := \begin{pmatrix} \int_{k\tau}^{k\tau+\theta} \mathbf{T}(k\tau+\theta-s)Bd_1(s)ds + \varepsilon \int_0^\theta \mathbf{T}(s)Bdsf(k\tau; d_1, d_2, r) \\ B_cE_c[-Dd_1(k\tau) + r(k\tau) - d_2(k\tau)] \end{pmatrix},$$

where

$$f(k\tau; d_1, d_2, r) := -D_cDEd_1(k\tau) + ED_c[r(k\tau) - d_2(k\tau)].$$

By (4.35)–(4.37) and a routine calculation, we obtain

$$\begin{pmatrix} x_p(k\tau+\theta) \\ x_c(k+1) \end{pmatrix} = \Delta(\theta)\begin{pmatrix} x_p(k\tau) \\ x_c(k) \end{pmatrix} + R(k,\theta)\begin{pmatrix} d_1 \\ d_2 \\ r \end{pmatrix} \quad \forall k \in \mathbb{Z}_+,\ \theta \in [0,\tau). \tag{4.40}$$

It follows from (4.40) with $\theta = \tau$ that

$$\begin{pmatrix} x_p((k+1)\tau) \\ x_c(k+1) \end{pmatrix} = \Delta(\tau)\begin{pmatrix} x_p(k\tau) \\ x_c(k) \end{pmatrix} + R(k,\tau)\begin{pmatrix} d_1 \\ d_2 \\ r \end{pmatrix} \quad \forall k \in \mathbb{Z}_+. \tag{4.41}$$

In the following, let $\tau \in (0,\tau^*)$. Applying statement 1 of Theorem 3.5 to the feedback interconnection of discrete-time systems (4.38) and (4.36), we conclude that there exists $\varepsilon_\tau > 0$ such that, for every $\varepsilon \in (0,\varepsilon_\tau)$, $\Delta(\tau)$ is power stable.

By the admissibility of B, there exists $M_1 \geq 0$ such that

$$\left\| \int_{k\tau}^{k\tau+\theta} \mathbf{T}(k\tau+\theta-s)Bd_1(s)ds \right\|_X = M_1\|d_1\|_{L^2((k\tau,k\tau+\theta),\mathbb{C}^m)} \leq M_1\sqrt{\tau}\|d_1\|_\infty$$
$$\forall k \in \mathbb{Z}_+,\ \forall \theta \in [0,\tau],\ \forall d_1 \in L_b(\mathbb{R}_+,\mathbb{C}^m).$$

Therefore, there exists $M_2 \geq 0$ such that

$$\left\| R(k,\theta)\begin{pmatrix} d_1 \\ d_2 \\ r \end{pmatrix} \right\| \leq M_2(\|r\|_\infty + \|d_1\|_\infty + \|d_2\|_\infty)\ \forall k \in \mathbb{Z}_+,\ \forall \theta \in [0,\tau], \tag{4.42}$$
$$\forall r,\ d_2 \in L_b(\mathbb{R}_+,\mathbb{C}^p),\ \forall d_1 \in L_b(\mathbb{R}_+,\mathbb{C}^m).$$

Hence, it follows from the discrete-time variation-of-parameters formula, the power stability of $\Delta(\tau)$, (4.41), and (4.42) that there exist $M_3 \geq 1$ and $\rho \in (0,1)$ such that

$$\left\| \begin{pmatrix} x_p(k\tau) \\ x_c(k) \end{pmatrix} \right\| \leq M_3 \left(\rho^k \left\| \begin{pmatrix} x_p^0 \\ x_c^0 \end{pmatrix} \right\| + \|r\|_\infty + \|d_1\|_\infty + \|d_2\|_\infty \right)\ \forall k \in \mathbb{Z}_+,\ \forall x_p^0 \in X,$$
$$\forall x_c^0 \in \mathbb{C}^{Np+n_0},\ \forall r,\ d_2 \in L_b(\mathbb{R}_+,\mathbb{C}^p),\ \forall d_1 \in L_b(\mathbb{R}_+,\mathbb{C}^m). \tag{4.43}$$

Setting $M_4 := \max_{\theta\in[0,\tau]} \|\Delta(\theta)\|$, it follows from (4.40), (4.42), and (4.43) that, for all $\theta \in [0,\tau)$, $k \in \mathbb{Z}_+$, $x_p^0 \in X$, $x_c^0 \in \mathbb{C}^{Np+n_0}$, r, $d_2 \in L_b(\mathbb{R}_+,\mathbb{C}^p)$, and $d_1 \in L_b(\mathbb{R}_+,\mathbb{C}^m)$,

$$\begin{aligned}\left\|\begin{pmatrix} x_p(k\tau+\theta) \\ x_c(k+1)\end{pmatrix}\right\| &\le M_4 \left\|\begin{pmatrix} x_p(k\tau) \\ x_c(k)\end{pmatrix}\right\| + M_2(\|r\|_\infty + \|d_1\|_\infty + \|d_2\|_\infty) \\ &\le M_3 M_4 \rho^k \left\|\begin{pmatrix} x_p^0 \\ x_c^0\end{pmatrix}\right\| + (M_2 + M_3M_4)(\|r\|_\infty + \|d_1\|_\infty + \|d_2\|_\infty) \\ &\le N_1\left(e^{\beta(k\tau+\theta)}\left\|\begin{pmatrix} x_p^0 \\ x_c^0\end{pmatrix}\right\| + \|r\|_\infty + \|d_1\|_\infty + \|d_2\|_\infty\right),\end{aligned}$$

where $\beta := (\ln\rho)/\tau < 0$ and $N_1 := \max\{M_3M_4\rho^{-1}, M_2 + M_3M_4\}$. This completes the proof of statement 1.

To prove the approximate tracking and disturbance rejection result claimed in statement 2, note that, by exponential stability of (4.34) and boundedness of C, the impulse response of (4.34) is a $\mathbb{C}^{p\times m}$-valued Borel measure μ of the form $\mu(ds) = g(s)ds + D\delta_0(ds)$, where $g(\cdot)e^{\alpha\cdot} \in L^1(\mathbb{R}_+,\mathbb{C}^{p\times m})$ for some $\alpha > 0$, and δ_0 is the Dirac measure (see [8, Lemma 2.3]). By (4.34)–(4.37) and a routine calculation, we obtain

$$(4.44)\qquad \begin{pmatrix} y(k\tau+\theta) \\ y_c(k)\end{pmatrix} = Q(\theta)\Delta^k(\tau)\begin{pmatrix} x_p^0 \\ x_c^0\end{pmatrix} + \begin{pmatrix} \tilde{y}(k\tau+\theta) \\ \tilde{y}_c(k)\end{pmatrix} \quad \forall\theta\in[0,\tau),\ \forall k\in\mathbb{Z}_+,$$

where

$$Q(\theta) := \begin{pmatrix} C\mathbf{T}(\theta) - \varepsilon(F(\theta)+DE)D_cC & \varepsilon F(\theta)C_c + \varepsilon DEC_c \\ -\varepsilon D_cE_cC & \varepsilon C_c - \varepsilon^2 D_cE_cDC_c\end{pmatrix},$$

$$\text{with } F(\theta) := C\textstyle\int_0^\theta \mathbf{T}(s)BdsE,$$

and $\tilde{y}$, $\tilde{y}_c$ satisfy

$$(4.45)\qquad \tilde{y} = G(d_1 + \mathcal{H}_\tau\tilde{y}_c) + d_2, \quad \tilde{y}_c = K_{\tau,\varepsilon}\mathcal{S}_\tau(r-\tilde{y}).$$

An application of Theorem 4.10 to system (4.45), with r given by (4.9) and d_1, d_2 satisfying (4.10), shows that for every $\delta > 0$, there exists $\tau_\delta \in (0,\tau^*)$ such that, for every sampling period $\tau \in (0,\tau_\delta)$, there exists $\varepsilon_\tau > 0$, such that, for every $\varepsilon \in (0,\varepsilon_\tau)$,

$$\limsup_{t\to\infty}\|\tilde{y}(t) - r(t)\| \le \delta.$$

Therefore, by power stability of $\Delta(\tau)$ and (4.44),

$$\limsup_{t\to\infty}\|y(t) - r(t)\| \le \delta \quad \forall x_p^0 \in X,\ \forall x_c^0 \in \mathbb{C}^{Np+n_0},$$

completing the proof. □

Example 4.13. For purposes of illustration, we consider the heat equation for a bar of length 1. We keep both endpoints at zero temperature and inject heat of magnitude u_p at the point $\eta_1 \in (0,1)$. The measurement is generated by a spatial averaging of the state over an σ-neighborhood of a point $\eta_2 \in (\eta_1, 1)$. The system to be controlled can be formulated as follows:

$$\begin{aligned} z_t(\eta,t) &= z_{\eta\eta}(\eta,t) + \delta(\eta-\eta_1)u_p(t), \\ y_p(t) &= \frac{1}{2\sigma}\int_{\eta_2-\sigma}^{\eta_2+\sigma} z(\lambda,t)d\lambda,\end{aligned}$$

with boundary conditions

$$z(0,t) = z(1,t) = 0 \quad \forall t > 0\,.$$

For simplicity, we assume zero initial conditions

$$z(\eta,0) = 0 \quad \forall \eta \in [0,1]\,.$$

Sampled-data low-gain integral control of this system (in the presence of input hysteresis) was studied in [6].

With input u_p and output y_p, it is not hard to show that this system is a regular linear system with state space $X = L^2(0,1)$ and bounded observation. In particular, the corresponding semigroup $\mathbf{T}(t)$, given by

$$(\mathbf{T}(t)x)(\eta) = \sum_{n=1}^{\infty} 2\exp(-n^2\pi^2 t)\sin(n\pi\eta)\int_0^1 \sin(n\pi\lambda)x(\lambda)d\lambda \qquad \forall x \in L^2(0,1),\ \ \forall \eta \in [0,1]\,,$$

is exponentially stable. The transfer function $\mathbf{G}$ is given by

$$\mathbf{G}(s) = \frac{\sinh(\sigma\sqrt{s})\sinh(\eta_1\sqrt{s})\sinh((1-\eta_2)\sqrt{s})}{\sigma s\sinh(\sqrt{s})}\,.$$

The aim is to design a robust controller such that the closed-loop system approximately tracks the reference signal $r(t) = \sin t$ in the presence of disturbance signals d_1, d_2 given by

$$d_1(t) = \frac{1}{5}\cos(5t) + \frac{1}{t+1}\,, \quad d_2(t) = \frac{1}{5}\sin(5t) - \frac{1}{2}\ln\left(1 + \frac{1}{t+1}\right)\,, \quad t \geq 0\,.$$

Set

$$K_1 := 1/\mathbf{G}(i)\,, \quad K_2 := \overline{K_1}\,, \quad K_3 := 1/\mathbf{G}(5i)\,, \quad K_4 := \overline{K_3}\,,$$

and $\mathbf{K}^0(z) \equiv 10$, so that the transfer function $\mathbf{K}_{\tau,\varepsilon}$ of the controller $K_{\tau,\varepsilon}$ (see (4.7)) is given by

$$\begin{aligned}\mathbf{K}_{\tau,\varepsilon}(z) &:= \varepsilon\left(10 + \frac{K_1}{z - e^{i\tau}} + \frac{K_2}{z - e^{-i\tau}} + \frac{K_3}{z - e^{5i\tau}} + \frac{K_4}{z - e^{-5i\tau}}\right)\\ &= \varepsilon\left(10 + \frac{2\mathrm{Re}\,(K_1)z - 2\mathrm{Re}\,(K_1e^{-i\tau})}{z^2 - (2\cos\tau)z + 1} + \frac{2\mathrm{Re}\,(K_3)z - 2\mathrm{Re}\,(K_3e^{-5i\tau})}{z^2 - (2\cos 5\tau)z + 1}\right)\,.\end{aligned}$$

Since all the relevant hypotheses are satisfied, the conclusions of Theorem 4.10 are valid. In Figure 4.2, simulations are shown for the specific values

$$\eta_1 = 0.2\,, \quad \eta_2 = 0.6\,, \quad \sigma = 0.01\,, \quad \tau = 0.1\,, \quad \varepsilon = 0.1\,,$$

with zero initial conditions for the controller. The error signal $e = r - y_p - d_2$ and the output of the sampled-data system $y = y_p + d_2$ are shown in Figure 4.2. Asymptotically, the error is bounded by 0.0028, that is, $\limsup_{t\geq 0}|e(t)| \leq 0.0028$. Simulations show that, for the sampling period $\tau = 0.1$, instability occurs at $\varepsilon \approx 0.22$.

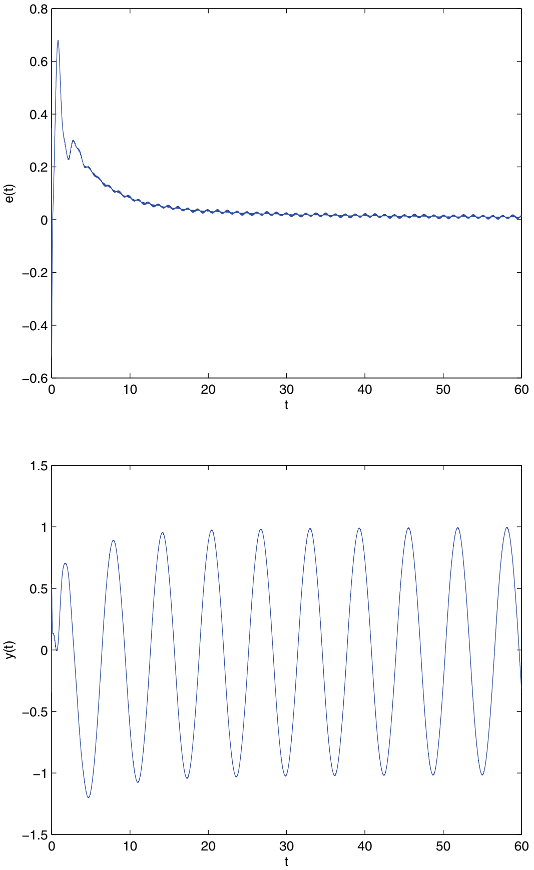


Fig. 4.2. *Error signal e and output y.*

REFERENCES

[1] V. H. L. Cheng and C. A. Desoer, *Discrete time convolution control systems*, Internat. J. Control, 36 (1982), pp. 367–407.

[2] E. J. Davison, *Multivariable tuning regulators: The feedforward and robust control of a general servomechanism problem*, IEEE Trans. Automat. Control, 21 (1976), pp. 35–47.

[3] T. Hämäläinen and S. Pohjolainen, *A finite-dimensional robust controller for systems in the CD-algebra*, IEEE Trans. Automat. Control, 45 (2000), pp. 421–431.

[4] Z. Ke, *Sampled-Data Control: Stabilization, Tracking and Disturbance Rejection*, Ph.D. thesis, University of Bath, 2008; available electronically from http://www.maths.bath.ac.uk/∼hl/THESES/ke_thesis.pdf.

[5] H. Logemann, *Stability and stabilizability of linear infinite-dimensional discrete-time systems*, IMA J. Math. Control Inform., 9 (1992), pp. 252–263.

[6] H. Logemann and A. D. Mawby, *Discrete-time and sampled-data low-gain control of infinite-dimensional linear systems in the presence of input hysteresis*, SIAM J. Control Optim., 41 (2002), pp. 113–140.

[7] H. Logemann and D. H. Owens, *Low-gain control of unknown infinite-dimensional systems: A frequency-domain approach*, Dynamics Stability Syst., 4 (1989), pp. 13–29.

[8] H. Logemann and E. P. Ryan, *Time-varying and adaptive integral control of infinite-dimensional regular linear systems with input nonlinearities*, SIAM J. Control Optim., 38 (2000), pp. 1120–1144.

[9] H. Logemann and S. Townley, *Low-gain control of uncertain regular linear systems*, SIAM J. Control Optim., 35 (1997), pp. 78–116.

[10] H. Logemann and S. Townley, *Discrete-time low-gain control of uncertain infinite-dimensional systems*, IEEE Trans. Automat. Control, 42 (1997), pp. 22–37.

[11] M. Morari, *Robust stability of systems with integral control*, IEEE Trans. Automat. Control, 30 (1985), pp. 574–577.

[12] S. Pohjolainen, *Robust multivariable PI-controller for infinite-dimensional systems*, IEEE Trans. Automat. Control, 27 (1982), pp. 17–30.

[13] R. Rebarber and G. Weiss, *Internal model based tracking and disturbance rejection for stable well-posed systems*, Automatica, 39 (2003), pp. 1555–1569.

[14] M. C. Smith, *On stabilization and the existence of coprime factorizations*, IEEE Trans. Automat. Control, 34 (1989), pp. 1005–1007.

[15] O. J. Staffans, *Well-Posed Linear Systems*, Cambridge University Press, Cambridge, UK, 2005.

[16] M. Vidyasagar, *Control System Synthesis*, MIT Press, Cambridge, MA, 1985

[17] M. Vidyasagar, H. Schneider, and B. A. Francis, *Algebraic and topological aspects of feedback stabilization*, IEEE Trans. Automat. Control, 27 (1982), pp. 880–894.

[18] G. Weiss, *Transfer functions of regular linear systems—part* I: *Characterization of regularity*, Trans. Amer. Math. Soc., 342 (1994), pp. 827–854.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 2, pp. 672–687

© 2009 Society for Industrial and Applied Mathematics

OPTIMAL CONTROL IN FLUID MECHANICS BY FINITE ELEMENTS WITH SYMMETRIC STABILIZATION[*]

M. BRAACK[†]

Abstract. There are two main possibilities for the numerical computation of optimal control problems with constraints given by partial differential equations: One may consider first the discretized problem and then build the optimality condition. The other possibility is to formulate first the optimality condition on the continuous level and then discretize. Both approaches may lead to different discrete adjoint equations because discretization and building the adjoint do not commute in general. This type of inconsistency takes place when conventional stabilized finite elements for flow problems, as for instance, streamline diffusion (SUPG), are used, due to its nonsymmetry. Consequently, the computed control is significantly affected by the way of defining the discrete optimality condition. Hence, there is a need for symmetric stabilization so that discretization and building the adjoint commute. We formulate the use of this kind of stabilization and give a quasi-optimal a priori estimate in the context of optimal control problems for the Oseen system. In particular, we show that local projection stabilization and edge-oriented stabilization result to be quasi-optimal for optimal control problems.

Key words. finite elements, stabilization, optimal control, Oseen equation, Navier–Stokes

AMS subject classifications. 35Q30, 65N12, 65N30, 76D05, 76D55, 76M10, 93C20

DOI. 10.1137/060653494

1. Introduction.

The numerical computation of optimal control problems with constraints given by partial differential equations can be divided into two main approaches: One may consider first the discretized problem and then build the optimality condition. The other possibility is to formulate first the optimality condition on the continuous level and then discretize. Both approaches lead to different discrete adjoint equations when discretization and building the adjoint do not commute. This type of inconsistency takes place when conventional residual-based stabilized finite elements for flow problems, as for instance *streamline upwind/Petrov–Galerkin* (SUPG) introduced by Brooks and Hughes in [9], are used, because they are nonsymmetric. Consequently, the computed control is significantly affected by the way of defining the discrete optimality condition.

An error estimate for convection-diffusion-reaction equations with the SUPG method is given by Collis and Heinkenschloss in [11] where the two approaches "discretize-optimize" and "optimize-discretize" different a priori estimates are derived. The estimate for "optimize-discretize" has a better asymptotic in terms of powers of mesh size. In numerical tests, the largest difference is observed in the adjoint variable. For convection-diffusion problems with a particular least-squares stabilization Dedé and Quarteroni [12] derived an a posteriori estimate and used it for local mesh refinement. Becker and Vexler [4] presented recently an a priori estimate for optimal control with such a scalar equation for finite elements with local projection stabilization.

Since more inconsistent terms appear in systems of equations, Abraham et al. investigated numerically the *Galerkin Least-Squares* (GLS) stabilization for the Oseen system in [1]. Herein, a significant effect is observed between both approaches.

[*] Received by the editors March 2, 2006; accepted for publication (in revised form) October 20, 2008; published electronically February 13, 2009.
http://www.siam.org/journals/sicon/48-2/65349.html

[†] Mathematisches Seminar, Christian-Albrechts-Universität Kiel, Ludewig-Meyn-Str. 4, 24098 Kiel, Germany (braack@math.uni-kiel.de).

Moreover, the computed control appears to be very sensitive to the evaluation of stabilization parameters. Therefore, the authors conclude that it is questionable whether the GLS approach is suitable for optimal control problems.

Li and Petzold [16] discussed this topic as well in the context of (a) consistent discrete boundary conditions for the adjoint problem and (b) adaptive mesh refinement. They propose a combination of "discretize-optimize" and "optimize-discretize" by splitting the domain into an inner part and a boundary part. The aspect of stabilization due to the presence of convective terms or due to a saddle point structure of the primal equation is not considered.

Obviously, there is a need for symmetric stabilization so that discretization and building the adjoint commute. Recently, new stabilization techniques which are not residual-based were developed and analyzed, for instance, edge-oriented stabilization (see Burman et al. [10]) or local projection stabilization (LPS) [2]. For a review and a critical comparison of those techniques with residual-based techniques, we refer to [6]. In this work we derive some sufficient conditions for stabilized finite elements in order to obtain a consistent and stable adjoint problem in the context of optimal control. Local projection and edge-oriented stabilization are two prototypical methods fulfilling these conditions.

As a prototype example we consider the variational formulation of the Oseen problem, which is a very popular linearization of the Navier–Stokes equations. In the domain $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, we consider the following system of equations for pressure p and velocity field v:

$$\begin{aligned} -\mu\Delta v + (\beta \cdot \nabla)v + \sigma v + \nabla p &= f \quad \text{in } \Omega\,, \\ \operatorname{div} v &= 0 \quad \text{in } \Omega\,, \\ v &= 0 \quad \text{on } \partial\Omega\,, \end{aligned} \tag{1.1}$$

with divergence-free convection field $\beta : \Omega \to \mathbb{R}^d$, $\operatorname{div}\beta = 0$, and viscosity parameter $\mu > 0$. The parameter $\sigma \geq 0$ may come from a possible discretization in time.

We explain now the principle dilemma arising in optimal control problems with discrete state equations: We write a linear state equation given by a partial differential equation in abstract operator form:

$$Au + Bq = f\,. \tag{1.2}$$

Here, u denotes the state variable, and q the control out of the subspace Q of $L^2(\Omega)$ with L^2-norm $\|\cdot\|$. For the Oseen system above, we use the vector notation $u := \{v, p\}$ in order to prevent confusion with the L^2-scalar product $(\cdot,\cdot)$. The operator A applied to the state vector u is given by the matrix-vector multiplication in function spaces:

$$Au := \begin{pmatrix} -\mu\Delta + (\beta\cdot\nabla) + \sigma & \nabla \\ \operatorname{div} v & 0 \end{pmatrix} \begin{pmatrix} v \\ p \end{pmatrix}.$$

The objective functional under consideration is of the form:

$$J(u, q) := \frac{1}{2}\|Cu - C\widehat{u}\|^2 + \frac{\alpha}{2}\|q\|^2\,, \tag{1.3}$$

with a regularization parameter $\alpha > 0$, a target state $\widehat{u}$. The operator C should be linear and L^2-continuous. Since tracking problems in fluid dynamics usually are focused on the velocities, we suppose

$$\|Cu\| \leq c\|v\| \tag{1.4}$$

with a positive constant c. Hence, the optimal control problem reads

$$\arg\min\Big\{J(u,q) : u \text{ is solution of (1.2) for control } q\in Q\Big\}.$$

The corresponding (continuous) Karush–Kuhn–Tucker (KKT) system is of the following form:

$$\begin{pmatrix}\alpha I & 0 & B^* \\ 0 & C & A^* \\ B & A & 0\end{pmatrix}\begin{pmatrix}q\\u\\z\end{pmatrix}=\begin{pmatrix}0\\C\widehat{u}\\f\end{pmatrix},$$

where $z=\{z^v,z^p\}$ denotes the adjoint state (z^v = adjoint velocity, z^p = adjoint pressure). If these three equations are discretized properly, we obtain a discrete KKT system. We provide the discrete variables with the subscribe h to emphasize the dependency of the mesh-size function h: u_h, z_h, q_h. If we denote the discrete operators with the subscript h as well, we obtain formally the corresponding discrete primal state equation:

$$A_h u_h + B_h q_h = f_h\,.$$

However, this discrete version is not necessarily stable, for instance, due to convective terms like in the Oseen system, $(\beta\cdot\nabla)v$. The principle idea of stabilized finite elements is the use of additional terms S_h^u, S_h^q taking account for such instabilities of the discrete operator A_h:

$$(A_h+S_h^u)u_h+(B_h+S_h^q)q_h=f_h\,. \tag{1.5}$$

In the case of the Oseen system, such term s may be necessary for satisfying the discrete inf-sup condition and to stabilize the indefinite convective terms. Since the adjoint equation has to be stabilized as well (let us denote this operator by S_h^z), the corresponding discrete KKT system of “optimize-discretize” is of the form

$$\begin{pmatrix}\alpha I & 0 & B_h^* \\ 0 & C_h & A_h^*+S_h^z \\ B_h+S_h^q & A_h+S_h^u & 0\end{pmatrix}\begin{pmatrix}q_h\\u_h\\z_h\end{pmatrix}=\begin{pmatrix}0\\C\widehat{u}\\f_h\end{pmatrix}.$$

The other possibility is to start with the discrete state (1.5), and build the corresponding KKT system (“discretize-optimize”), cf. [11]:

$$\begin{pmatrix}\alpha I & 0 & B_h^*+([S_h^q]')^* \\ 0 & C_h & A_h^*+([S_h^u]')^* \\ B_h+S_h^q & A_h+S_h^u & 0\end{pmatrix}\begin{pmatrix}q_h\\u_h\\z_h\end{pmatrix}=\begin{pmatrix}0\\C\widehat{u}\\f_h\end{pmatrix}.$$

Here, the additional term $([S_h^q]')^*$ arises in the equation for q_h, which is the adjoint linearization of S_h^q. These two strategies coincide only in the case that it holds $S_h^z = ([S_h^u]')^*$ and $S_h^q=0$, which is not true in general. We will come back to this point in section 3 for the established SUPG-PSPG (streamline upwind Petrov–Galerkin–pressure stabilized Petrov–Galerkin) stabilization of the Oseen system.

In this work, we derive sufficient conditions for stabilized finite elements so that the resulting discrete optimal control problem allows for a quasi-optimal a priori estimate. Particular examples of those methods are local projection stabilization (LPS) and edge-oriented stabilization (EOS). We show that the Oseen system stabilized with

LPS or EOS leads to a consistent discrete KKT system in the sense that "discretize-optimize" and "optimize-discretize" coincide. We give on quasi-uniform triangulations with maximum mesh size h an a priori error estimate of the form:

$$\|\!|u - u_h|\!\|_h + \|q - q_h\| \lesssim h^{r+\frac{1}{2}} (\|u\|_{r+1} + \|z\|_{r+1} + \|q\|_{r+1})\,, \tag{1.6}$$

for a certain h-dependent (semi) norm $\|\!|\cdot|\!\|_h$. Here r stands for the polynomial degree of the finite elements and $\|\cdot\|_{r+1}$ for the norm in the Sobolev space $H^{r+1}(\Omega)$. The expression $a \lesssim b$ means $a \le cb$ for a generic h-independent constant c.

The contents of this work are as follows: In the following section, we formulate the variational formulation of the Oseen system and derive the necessary and sufficient condition for an optimal control problem on the continuous level. The discrete KKT system is formulated for finite element discretizations. In particular, the approaches *discretize-optimize* and *optimize-discretize* for stabilized finite element discretization will be presented. In section 3 we derive conditions on the stabilization, namely, symmetry, coercivity, and quasi-optimality for the primal problem with given control. For the resulting discrete optimal control problems, *discretize-optimize* and *optimize-discretize* coincides, and an a priori estimate of similar quality as the solely primal problem (for given control) is proven, i.e., (1.6). In section 4 we show that two classes of stabilization techniques, LPS, and EOS, fit into this class of finite element techniques. Finally, we extend LPS to the Navier–Stokes system and discuss shortly the consequences for the corresponding KKT system.

2. Variational formulation for optimal control.

2.1. Continuous variational formulation. The state variable $u : \Omega \to \mathbb{R}^m$ is sought in a Banach space $X \subset [L^2(\Omega)]^m$ and the control space Q is a subspace of L^2-integrable functions, $Q \subset L^2(\Omega)$. For the Oseen system we have $m = d + 1$. The equation for the state is given in the form

$$a(u, \varphi) + b(q, \varphi) = \langle l, \varphi\rangle \quad \forall \varphi \in X, \tag{2.1}$$

with bilinear forms $a : X \times X \to \mathbb{R}$ and $b : Q \times X \to \mathbb{R}$. A possible control is, e.g., $b(q, \varphi) = (Bq, \varphi)$ with a continuous linear mapping $B : Q \to [L^2(\Omega)]^m$. The right-hand side is a functional $l \in (L^2(\Omega)^m)'$. For the Oseen system (1.1) this functional acts as L^2-integral in the momentum equation, $\langle l, \varphi\rangle = \int_\Omega f \varphi_v \, dx$. As a typical optimal control problem one may consider a functional J as in (1.3). With a linear operator $C : X \to [L^2(\Omega)]^m$ the difference between Cu and a target solution $C\widehat{u} \in X$ in the $L^2(\Omega)$-norm is measured, together with the costs in the L^2-norm of the control weighted by some positive constant $\alpha > 0$.

We assume that the bilinear forms a and b are such that for every control $q \in Q$ there exists a unique solution $u \in X$ of (2.1) and

$$u = \mathcal{S}(q) \equiv u_0 + Tq\,, \tag{2.2}$$

where u_0 is a solution of (2.1) with $q = 0$ and $T : Q \to X$ is a bounded linear operator. Hence, the optimization problem can be expressed as

$$\{u, q\} = \arg\min\{J(u, q) : q \in Q,\, u = \mathcal{S}(q)\}\,. \tag{2.3}$$

2.2. First-order continuous optimality condition. We begin with the first-order optimality system:

Lemma 2.1. *A necessary and sufficient condition for a solution $\{u,q\}$ of the optimal control problem* (2.3) *is the existence of a dual state $z \in X$, so that the following system is fulfilled:*

$$a(u,\varphi) + b(q,\varphi) - \langle l,\varphi\rangle = 0 \qquad \forall\varphi \in X\,, \tag{2.4}$$

$$a(\psi,z) + (Cu - C\widehat{u}, C\psi) = 0 \qquad \forall\psi \in X\,, \tag{2.5}$$

$$\alpha(q,\lambda)_Q + b(\lambda,z) = 0 \qquad \forall\lambda \in Q\,. \tag{2.6}$$

Proof. A more general proof of this result can be found in the book of Lions [17]. However, we recall this proof and make it more specific since we need several parts of this proof later.

In order to show that the system (2.4)–(2.6) is a necessary condition, we can suppose that $\{u,q\}$ is given by (2.3). The state equation (2.4) is fulfilled by the definition of $\mathcal{S}$. Furthermore, we introduce the adjoint state $z \in X$ as the solution of (2.5). Hence, it remains to show the validity of (2.6). Due to (2.2) the reduced functional $j : Q \to \mathbb{R}$, $j(q) := J(\mathcal{S}(q),q)$ has the derivative:

$$\begin{aligned} j'(q)(\lambda) &= (Cu - C\widehat{u}, C\mathcal{S}\,'(q)\lambda) + \alpha(q,\lambda)_Q \\ &= (Cu - C\widehat{u}, CT\lambda) + \alpha(q,\lambda)_Q\,. \end{aligned}$$

Since the necessary condition to minimize j is $j'(q)(\lambda) = 0$ for all $\lambda \in Q$, we obtain for the optimal control q:

$$\begin{aligned} \alpha(q,\lambda)_Q &= -(Cu - C\widehat{u}, CT\lambda) \\ &= (Cu - C\widehat{u}, C(u_0 - \mathcal{S}(\lambda)))\,. \end{aligned}$$

Using the adjoint state $z \in X$ given by (2.5), we get for $\psi := u_0 - \mathcal{S}(\lambda)$:

$$a(u_0 - \mathcal{S}(\lambda), z) = -(Cu - C\widehat{u}, C(u_0 - \mathcal{S}(\lambda)))\,.$$

By combining the last two equations, we get due to linearity:

$$\begin{aligned} \alpha(q,\lambda)_Q &= a(\mathcal{S}(\lambda) - u_0, z) \\ &= a(\mathcal{S}(\lambda), z) - \langle l,z\rangle \\ &= \langle l,z\rangle - b(\lambda,z) - \langle l,z\rangle\,. \end{aligned}$$

This proves the necessary condition. In order to show that the system of equations is a sufficient condition, we consider the second derivative of the reduced functional:

$$j''(q)(\lambda,\lambda) = \|CT\lambda\|^2 + \alpha\|\lambda\|^2 \;\geq\; 0\,. \tag{2.7}$$

Hence we get also the sufficient condition. □

2.3. Stabilized finite elements. In order to solve the optimality system (2.4)–(2.6) numerically, we will replace the infinite dimensional space X by a finite dimensional subspace $X_h \subset X$ consisting of conforming finite elements. However, since the Galerkin system is not necessarily stable, one has to add certain stabilization terms. This will be made more concrete in this section. After discretization of the system, one may solve the corresponding finite dimensional problem totally coupled or iteratively, for instance, by a gradient method.

The mesh size function will be denoted by h and the triangulation by $\mathcal{T}_h$. We consider shape regular meshes $\mathcal{T}_h$ of tetrahedral or hexahedral elements K in three spatial dimensions ($d = 3$), and triangular or quadrilateral elements for the two-dimensional case ($d = 2$). The finite element spaces result from isoparametric transformations of polynomials on a reference cell $\hat{K}$. By $\mathcal{P}^r$ we denote the set of polynomials of total degree (for hex's/quad's) or rather maximal degree (for tri's/tet's) r on the reference cell $\hat{K}$, and $T_K : \hat{K} \to K$ a polynomial transformation of the same type and degree; i.e., $T_K \in \mathcal{P}^r$. Now, we formulate the finite element space

$$\mathcal{P}_h^r := \left\{ \varphi \in C(\Omega, \mathbb{R}) : \varphi|_K = \hat{\varphi} \circ T_K^{-1} \text{ with } \hat{\varphi}, T_K \in \mathcal{P}^r \right\} .$$

These finite element spaces are the usual P_r elements (tri's/tet's) or rather Q_r elements (quad's/hex's). The discrete space is the intersection of X and the product space of these finite elements (equal-order finite elements):

$$X_h = X \cap [\mathcal{P}_h^r]^m .$$

The control space is also approximated by a certain discrete space $Q_h \subset Q$. We may take, for instance, the same type of finite elements as for u, but this is not mandatory.

In terms of stabilized finite elements the discrete state equation reads

$$(2.8) \qquad u_h \in X_h : \quad a(u_h, \varphi) + b(q_h, \varphi) + s_h^u(u_h, \cdot; \varphi) = \langle l, \varphi \rangle \quad \forall \varphi \in X_h .$$

The subscribe h in the stabilization term $s_h^u(u_h, \cdot; \varphi)$ indicates the dependence on the mesh-size, and the capital u indicates that it is the stabilization for the primal variable u. The stabilization term may depend on further quantities in the primal equation (we indicated this by the "$\cdot$"), as, for instance, the right-hand side f and on the control q_h as well. This holds, in particular, for conventional residual-based stabilization techniques.

The stabilization has to be chosen in such a way that there is also a unique discrete solution operator $\mathcal{S}_h : Q_h \to X_h$, $u_h = \mathcal{S}_h(q_h)$. For the adjoint equation (2.5) and the gradient equation (2.6) we may distinguish between *discretize-optimize* and *optimize-discretize* as discussed in the following.

2.3.1. Discretize-optimize. The approach *discretize-optimize* means that we start with a discrete primal equation (2.8) and formulate on its basis the corresponding discrete optimality conditions. Since the dual equation is always linear, we have to linearize the probably nonlinear stabilization term. By $\partial_u s_h^u(\cdot)(\varphi, \dots)$ we denote the Gateaux derivative with respect to u in the direction of φ:

$$\partial_u s_h^u(u, \cdot)(\varphi, z) := \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left(s_h^u(u + \epsilon\varphi, \cdot; z) - s_h^u(u, \cdot; z) \right) .$$

We obtain the adjoint equation:

$$(2.9) \quad z_h \in X_h : \quad a(\varphi, z_h) + \partial_u s_h^u(u_h, \cdot)(\varphi, z_h) = (C(\widehat{u} - u_h), C\varphi) \quad \forall \varphi \in X_h.$$

Although (2.8) is stable the adjoint equation (2.9) may not lead to the optimal convergence order. This is, indeed, the case for conventional stabilizations of flow problems, as, for instance, PSPG and SUPG. The reason for the suboptimality is that the term $\partial_u s_h^u(\cdot, \cdot)$ does not contain the full adjoint residual corresponding to the adjoint momentum equation arising in (2.5).

Furthermore, as shown in [11], the derivative of the stabilization with respect to the control enters into the discrete version of (2.6) in the case "discretize-optimize":

$$(2.10) \qquad q_h \in Q_h : \ \alpha(q_h, \lambda)_Q + b(\lambda, z_h) + \partial_q s_h^u(u_h, \cdot)(\lambda, z_h) = 0 \quad \forall \lambda \in Q_h .$$

The resulting discrete KKT system (2.8), (2.9), and (2.10) is fully consistent in the sense that for the discrete solution $\{u_h, q_h, z_h\}$ holds:

$$\{u_h, q_h\} = \arg\min\{J(u_h, q_h) : q_h \in Q,\, u_h = \mathcal{S}_h(q_h)\}\,. \tag{2.11}$$

2.3.2. Optimize-discretize. Alternatively, one may start with the continuous dual problem (2.5) and formulate a corresponding stable discrete version. Hence, this alternative starts firstly with the continuous optimality system and secondly discretizes properly by adding appropriate stabilization terms. Applying similar stabilization techniques as for the primal problem (2.4) leads to a term $s_h^z(\cdot,\cdot;\cdot)$ in the discrete adjoint equation:

$$z_h \in X_h : \quad a(\varphi, z_h) + s_h^z(z_h, \cdot; \varphi) = (C(\widehat{u} - u_h), C\varphi) \qquad \forall \varphi \in X_h\,. \tag{2.12}$$

In residual-based stabilization techniques, the stabilization involves the residual of the adjoint equation and is, therefore, dependent of $C(\widehat{u} - u_h)$. The discrete version of the gradient equation (2.6) is simply

$$q_h \in Q_h : \quad \alpha(q_h, \lambda) + b(\lambda, z_h) = 0 \qquad \forall \lambda \in Q_h\,. \tag{2.13}$$

Depending on the specific stabilization, the system (2.8), (2.12), and (2.13) is not a necessary condition to minimize J in $X_h \times Q_h$. Therefore, we obtain an inconsistent discrete KKT system. In summary, we see that there is a certain conflict between a consistent "suboptimal" and an inconsistent but "optimal" discrete adjoint problem. With "(sub)optimal" we refer to the discretization error in dependence of the mesh size.

3. Symmetric stabilization for optimal control. We formulate the abstract setting of the previous sections for the Oseen system (1.1). The state variable u consists of velocity $v \in V := H_0^1(\Omega)$ and pressure $p \in P := L_0^2(\Omega)$, the space of L^2-integrable function in Ω with zero mean. The bilinear form for the Oseen system with $u = \{v, p\}$ and test functions $\varphi = \{\phi, \xi\}$ reads:

$$a(u, \varphi) := (\operatorname{div} v, \xi) + (\sigma v, \phi) + (\beta \cdot \nabla v, \phi) + (\mu \nabla v, \nabla \phi) - (p, \operatorname{div} \phi)\,. \tag{3.1}$$

We write the control in the form $b(q, \varphi) = (Bq, \varphi)$.

The PSPG was introduced by Hughes et al. in [14]. It is introduced to circumvent the inf-sup condition. SUPG stabilization was developed independently by Johnson and Saranen in [15] and Brooks and Hughes [9] in order to stabilize the convective terms. The combination of PSPG and SUPG is a standard method for equal-order finite elements in computational fluid dynamics. This combination for the Oseen system with equal-order finite elements uses the stabilization term:

$$s_h^u(u, f - Bq; \varphi) := (\sigma v + (\beta \cdot \nabla) v + \nabla p + Bq - f, \delta^p \nabla \xi + \delta^v (\beta \cdot \nabla) \phi) \tag{3.2}$$

$$- \sum_K (\mu \Delta v, \delta^p \nabla \xi + \delta^v (\beta \cdot \nabla) \phi)_K\,, \tag{3.3}$$

where the parameters δ^p and δ^v are cell-wise constants, depending on the local Peclet number. It is easy to check that for this method in general holds $\partial_u s_h^u(u_h, \cdot)(\varphi, z_h) \not\equiv s_h^z(z_h, \cdot; \varphi)$ and $\partial_q s_h^u(u_h, \cdot) \not\equiv 0$.

Now we derive an a priori estimate for the optimal control problem for a quite general class of stabilized finite element schemes. The requirements we need consist in

- linearity and symmetry of the stabilization,
- a coercivity property, and
- quasi-optimality of the forward problem (i.e., for fixed control).

These properties are concretized in the following subsection.

3.1. Requirements on the finite element stabilization. From now on we focus on linear stabilization techniques for the Oseen system which are independent of the right-hand side f. For a given control $q \in Q$ such schemes are of the form

$$a(u_h,\varphi) + b(q,\varphi) + s_h(u_h,\varphi) = \langle l,\varphi\rangle \qquad \forall \varphi \in X_h\,. \tag{3.4}$$

For the bilinear stabilization form $s_h : X_h \times X_h \to \mathbb{R}$ we assume a symmetry property:

Lemma 3.1. *If a stabilized scheme of type* (3.4) *fulfills the symmetry property*

$$(P1) \qquad s_h(u,\varphi) = s_h(\varphi,u) \qquad \forall u,\varphi \in X\,,$$

then discretization and optimization commute.

Proof. Since $\partial_q s(u_h,\varphi) = 0$ the KKT system corresponding to the discretized primal equation reads:

$$a(u_h,\varphi) + b(q_h,\varphi) + s_h(u_h,\varphi) - \langle l,\varphi\rangle = 0 \qquad \forall \varphi \in X_h \tag{3.5}$$

$$a(\psi,z_h) + s_h(\psi,z_h) + (C(u_h-\widehat{u}),C\psi) = 0 \qquad \forall \psi \in X_h \tag{3.6}$$

$$\alpha(q_h,\lambda) + b(\lambda,z_h) = 0 \qquad \forall \lambda \in Q_h\,. \tag{3.7}$$

If we build the continuous KKT system first and then discretize, we obtain the same system because the stabilization term for (2.5) is $s_h(z_h,\psi)$. Due to (P1) this is equivalent to the discrete adjoint equation above. □

The a priori error estimate we are going to prove is in terms of a (possible h-dependent) semi-norm

$$|\!|\!| \cdot |\!|\!|_h : X \to \mathbb{R}_0^+\,,$$

which is supposed to fulfill the following (discrete) coercivity property:

$$(P2): \qquad |\!|\!| u_h |\!|\!|_h \lesssim (a(u_h,u_h) + s_h(u_h,u_h))^{1/2} \qquad \forall u_h \in X_h\,. \tag{3.8}$$

This semi-norm is supposed to be stronger than the L^2-norm of the velocities, i.e.,

$$(P3): \qquad \|v\| \lesssim |\!|\!| u |\!|\!|_h \qquad \forall u = \{v,p\} \in X\,. \tag{3.9}$$

We introduce the term "symmetric of order s" for a stabilized scheme with finite elements of order r.

Definition 3.2. *A stabilized scheme* (3.4) *with property (P1) is called* symmetric of order s, *with* $0 \le s \le r+1$, *if for (arbitrary but fixed) control* $q \in Q \subset H^{r+1}(\Omega)$ *the solution* $u_h = u_h(q)$ *of the discrete problem* (3.4) *and the solution* $u = u(q) \in [H^{r+1}(\Omega)]^{d+1}$ *of the continuous problem* (2.1) *fulfill the following a priori estimate:*

$$(P4) \qquad |\!|\!| u(q) - u_h(q) |\!|\!|_h \lesssim h^s \|u\|_{r+1} \tag{3.10}$$

for a triple norm satisfying conditions (P2) and (P3).

In this definition we assume the same regularity for p as for v. However, this condition can easily be relaxed to one order less of regularity for p; see [5].

3.2. A priori estimates. In this part we derive, firstly, a L^2-error estimate of the control, and secondly, an estimate for the error in the primal state and dual state.

For $w \in X$, we denote by $z(w)$ and $z_h(w)$ the continuous and discrete solutions of the adjoint problems, respectively, with a right-hand side including w:

$$\begin{aligned} z(w) \in X : \quad & a(\psi, z) &&= (C(\widehat{u} - w), C\psi) \qquad \forall \psi \in X\,, \\ z_h(w) \in X_h : \quad & a(\psi, z_h) + s_h(\psi, z_h) &&= (C(\widehat{u} - w), C\psi) \qquad \forall \psi \in X_h\,. \end{aligned}$$

LEMMA 3.3. *We consider a stabilized scheme which is symmetric of order s for the primal equation in the sense of Definition* 3.2 *and apply this to the discrete adjoint problem. Then it holds for arbitrary $w \in V$:*

$$\|\!|z(w) - z_h(w)|\!\|_h \lesssim h^s \|z(w)\|_{r+1}\,.$$

Proof. The adjoint problem is also of Oseen type. Hence, the estimate of type (3.10) carries over to the corresponding adjoint solutions. ∎

LEMMA 3.4. *For the differences of the discrete velocities $v_h(q + \delta q)$ and $v_h(q)$ for the control q and $q + \delta q$, respectively, as well as for the dual solutions $z_h(u)$ and $z_h(u + \delta u)$ for the data u and $u + \delta u = \{v + \delta v, p + \delta p\}$, respectively, it holds for their velocity components:*

$$\begin{aligned} \|v_h(q + \delta q) - v_h(q)\| &\lesssim \|\delta q\|\,, \\ \|z_h^v(u + \delta u) - z_h^v(u)\| &\lesssim \|\delta v\|\,. \end{aligned}$$

Proof. With the coercivity (3.8) and the continuity of the linear form b we obtain for $\delta v_h = v_h(q + \delta q) - v_h(q)$:

$$\begin{aligned} \|\delta v_h\|^2 \lesssim \|\!|\delta u_h|\!\|_h^2 &\lesssim a(\delta u_h, \delta u_h) + s_h(\delta u_h, \delta u_h) \\ &= \langle l, \delta u_h\rangle - b(q + \delta q, \delta v_h) - (\langle l, \delta u_h\rangle - b(q, \delta v_h)) \\ &= -b(\delta q, \delta v_h) \lesssim \|\delta q\| \|\delta v_h\|\,. \end{aligned}$$

This is the first estimate of the Lemma. The second estimate for the difference in the adjoint solution, $\delta z_h := z_h(u + \delta u) - z_h(u)$, it follows analogously due to the continuity (1.4) of the linear operator C:

$$\|\delta z_h^v\|^2 \lesssim \|\!|\delta z_h|\!\|_h^2 \lesssim a(\delta z_h, \delta z_h) + s_h(\delta z_h, \delta z_h) = -(C\delta u, C\delta z_h) \ \lesssim\ \|\delta v\|\, \|\delta z_h^v\|\,.$$

This gives us the assertion. ∎

LEMMA 3.5. *We suppose $\{u, z, q\} \in [H^{r+1}(\Omega)]^{3d+2}$ for the continuous solution of the optimal control problems* (2.4)–(2.6). *The stabilized scheme is assumed to be symmetric of order $s \le r + 1$ in the sense of Definition* 3.2. *Then it holds for the control q_h of the discretized system* (3.5)–(3.7):

$$\|q - q_h\| \lesssim h^s (\|u\|_{r+1} + \|z\|_{r+1} + \|q\|_{r+1})\,. \tag{3.11}$$

Proof. We split the error in the control in interpolation error and projection error:

$$\|q - q_h\| \ \le\ \|q - i_h q\| + \|i_h q - q_h\|\,.$$

If we take as i_h the nodal interpolation, it holds $\|q - i_h q\| \lesssim h^{r+1}\|q\|_{r+1}$. Since $r + 1 \ge s$, it is sufficient to bound the projection part. Denoting the discrete reduced functional by

$$j_h(q) := J(\mathcal{S}_h(q), q)\,,$$

with the discrete solution operator $u_h(q) = \mathcal{S}_h(q)$, the reduced optimization problems are

$$\min_{q\in Q} j(q) \qquad \text{and} \qquad \min_{q_h\in Q_h} j_h(q_h)\,.$$

The corresponding (continuous and discrete) optimality conditions are

$$\begin{aligned} j'(q)(\delta q) &= b(\delta q, z^v) + (\alpha q, \delta q) = 0 \quad \forall \delta q \in Q\,,\\ j_h'(q_h)(\delta q) &= b(\delta q, z_h^v) + (\alpha q_h, \delta q) = 0 \quad \forall \delta q \in Q_h\,. \end{aligned}$$

From (2.7) it follows $j''(q)(\lambda,\lambda) \geq \alpha\|\lambda\|^2$ for all $\lambda \in Q$. By the same arguments it holds also for the discrete reduced functional:

$$j_h''(q_h)(\delta q, \delta q) \geq \alpha\|\delta q\|^2 \quad \forall \delta q \in Q_h\,.$$

Furthermore, since $j_h(q)$ is at most quadratic, it implies for arbitrary $\delta q_h \in Q_h$:

$$j_h''(q_h)(\cdot, \delta q_h) = j_h'(q_h + \delta q_h)(\cdot) - j_h'(q_h)(\cdot)\,.$$

Hence, for arbitrary $\delta q_h \in Q_h$:

$$\begin{aligned} \alpha\|\delta q_h\|^2 &\leq j_h''(q_h)(\delta q_h, \delta q_h)\\ &= j_h'(q_h + \delta q_h)(\delta q_h) - j_h'(q_h)(\delta q_h)\,. \end{aligned}$$

Let us denote the discrete solution of the adjoint equation with right-hand side $u_h(i_h q)$ by $\widehat{z}_h := z_h(u_h(i_h q))$. Due to the optimality conditions, $j_h'(q_h)(\delta q_h) = 0 = j'(q)(\delta q_h)$, it follows especially for $\delta q_h = i_h q - q_h$:

$$\begin{aligned} \alpha\|i_h q - q_h\|^2 &\leq j_h'(i_h q)(i_h q - q_h) - j'(q)(i_h q - q_h)\\ &= b(i_h q - q_h, \widehat{z}_h^v - z^v) + (\alpha(\,i_h q - q), i_h q - q_h)\\ &\leq c\|\widehat{z}_h^v - z^v\| \cdot \|i_h q - q_h\| + \alpha\|i_h q - q\| \cdot \|i_h q - q_h\|\,. \end{aligned}$$

In the last step we used the fact that $b(\cdot,\cdot)$ is continuous. Dividing both sides by $\alpha\|i_h q - q_h\|$ gives

$$\|i_h q - q_h\| \leq \frac{c}{\alpha}\|\widehat{z}_h^v - z^v\| + \|i_h q - q\|\,.$$

Hence, it remains to bound $\|\widehat{z}_h^v - z^v\|$ by the right-hand side of (3.11). This can be done by a further splitting:

$$\begin{aligned} \|\widehat{z}_h^v - z^v\| &= \|z_h^v(u_h(i_h q)) - z^v(u(q))\|\\ &\leq \|z_h^v(u_h(i_h q)) - z_h^v(u(q))\| + \|z_h^v(u(q)) - z^v(u(q))\|\,. \end{aligned}$$

The first term $\|z_h^v(u_h(i_h q)) - z_h^v(u(q))\|$ can be bounded by the stability property in Lemma 3.4 of the discrete primal and adjoint problems:

$$\begin{aligned} \|z_h^v(u_h(i_h q)) - z_h^v(u(q))\| &\lesssim \|v_h(i_h q) - v(q)\|\\ &\leq \|v_h(i_h q) - v_h(q)\| + \|v_h(q) - v(q)\|\\ &\lesssim \|i_h q - q\| + h^s\|u(q)\|_{r+1}\\ &\lesssim h^{r+1}\|q\|_{r+1} + h^s\|u(q)\|_{r+1}\\ &\lesssim h^s(\|u\|_{r+1} + \|q\|_{r+1})\,. \end{aligned}$$

Here, we used the stability for the primal equation corresponding to Lemma 3.4 for the adjoint state. The second term $\|z_h^v(u(q)) - z^v(u(q))\|$ (which is based on the same control q) can be treated by the approximation property of the discrete adjoint problem (Lemma 3.3) and (3.9):

$$\|z_h^v(u(q)) - z^v(u(q))\| \lesssim |\!|\!| z_h(u(q)) - z(u(q)) |\!|\!|_h \lesssim h^s \|u(q)\|_{r+1} . \qquad \square$$

The main result is given in the following theorem.

THEOREM 3.6. *The stabilized scheme is assumed to be symmetric of order s in the sense of Definition* 3.2 *with a bilinear form* s_h. *Then it holds the following a priori estimate for* $u = u(q)$ *and* $u_h = u_h(q_h)$*:*

$$|\!|\!| u - u_h |\!|\!|_h \lesssim h^{r+\frac{1}{2}} (\|u\|_{r+1} + \|z\|_{r+1} + \|q\|_{r+1}) , \tag{3.12}$$

presumed the regularity $\{u, z, q\} \in [H^{r+1}(\Omega)]^{3d+2}$ *for the continuous solution of the optimal control problem* (2.3).

Proof. In order to get this estimate we start with

$$|\!|\!| u - u_h |\!|\!|_h \le |\!|\!| u(q) - u_h(q) |\!|\!|_h + |\!|\!| u_h(q) - u_h(q_h) |\!|\!|_h .$$

Due to the quasi-optimality (3.10) for $|\!|\!| \cdot |\!|\!|$, the first term $|\!|\!| u(q) - u_h(q) |\!|\!|_h$ is bounded by the right-hand side of (3.12). The second term can be bounded due to stability of the discrete problem with respect to the control. At first, we note that due to (P2):

$$\begin{aligned} |\!|\!| u_h |\!|\!|_h^2 \lesssim a(u_h, u_h) + s_h(u_h, u_h) &= (f, v_h) - (Bq_h, v_h) \\ \le (\|f\| + \|Bq_h\|) \|v_h\| &\lesssim (\|f\| + \|Bq_h\|) |\!|\!| u_h |\!|\!|_h . \end{aligned}$$

Hence, the discrete solution depends continuously on the right-hand side f and the control q_h: $|\!|\!| u_h |\!|\!| \lesssim \|f\| + \|Bq_h\|$. Due to linearity, this implies that for discrete solutions to different controls it holds

$$|\!|\!| u_h(q) - u_h(q_h) |\!|\!|_h \lesssim \|B(q - q_h)\| \lesssim \|q - q_h\| .$$

The previously derived bound in the control (3.11) gives an adequate bound of this term. $\square$

Note that the same estimate can easily be derived analogously for $|\!|\!| z - z_h |\!|\!|_h$.

4. Application to symmetric stabilization techniques. In this part we present the symmetric stabilization techniques by (a) local projection [5] and by (b) interior penalties/edge stabilization [10] for the Oseen system which fulfill the assumptions of Theorem 3.6.

4.1. Local projection stabilization. The main idea of local projection stabilization is to include fine grid fluctuations of the pressure and of the velocity gradient in the stabilization term. A very particular feature is that the stabilization disappears for coarse grid test functions. This technique has shown its potential in several incompressible and compressible fluid dynamical applications; see, e.g., [7, 8]. In order to specify the stabilization term, we have to introduce further notations. The discontinuous analogon of $\mathcal{P}_h^r$ is denoted by

$$\mathcal{P}_{h,disc}^r := \left\{ \varphi \in L^2(\Omega, \mathbb{R}) : \varphi|_K = \hat\varphi \circ T_K^{-1} \text{ with } \hat\varphi, T_K \in \mathcal{P}^r \right\} .$$

Furthermore, let $\mathcal{T}_{2h}$ be the coarser mesh obtained by a "global coarsening" of $\mathcal{T}_h$. For quadrilaterals or hexahedrals, the finer mesh $\mathcal{T}_h$ contains 2^d times more elements

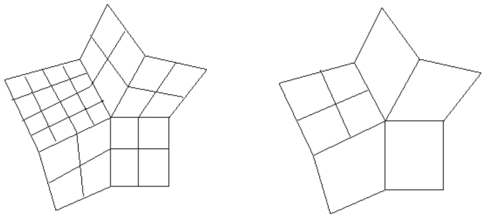


FIG. 4.1. *Patches of elements for the local projection stabilization: $\mathcal{T}_h$ (left) and $\mathcal{T}_{2h}$ (right).*

than $\mathcal{T}_{2h}$. The elements of $\mathcal{T}_{2h}$ will be denoted by "patches". In Figure 4.1 a typical mesh of this type is shown.

Let D_h^v and D_h^p be the following space for pressure and velocities, respectively, of functions allowing discontinuities across elements of $\mathcal{T}_{2h}$:

$$D_h^v := [\mathcal{P}^{r-1}_{2h,disc}]^d\,, \qquad D_h^p := \mathcal{P}^{r-1}_{2h,disc}\,.$$

In the case $r = 1$, these spaces contain patch-wise ($K \in \mathcal{T}_{2h}$) constants, and for $r = 2$, they contain patch-wise linear elements (d-linear elements for quad's or hex's). We make use of the L^2-projection operator

$$\bar{\pi}_h : L^2(\Omega) \to D_h^p\,,$$

characterized by the property for $v \in L^2(\Omega)$:

$$(v - \bar{\pi}_h v, \phi) = 0 \quad \forall \phi \in D_h^p\,.$$

Important is the fact that this projection acts locally on patches of elements, so that the numerical effort for computing this projection is very low. The operator giving the space fluctuations is denoted by

$$\bar{\varkappa}_h := i - \bar{\pi}_h\,,$$

with the identity mapping i. For ease of presentation, we use the same notations $\bar{\pi}_h$, $\bar{\varkappa}_h$ for the mappings on vector-valued functions, for instance, $\bar{\pi}_h : L^2(\Omega)^d \to D_h^v$.

The discrete primal equation of the optimal control problem with local projection is as in (3.4) with the stabilization term independent of the control q:

$$s_h^{lps}(u,\varphi) := (\beta\cdot\nabla v, \delta^v \bar{\varkappa}_h(\beta\cdot\nabla)\phi) + (\nabla p, \delta^p \bar{\varkappa}_h \nabla\xi) + (\operatorname{div} v, \gamma\bar{\varkappa}_h(\operatorname{div}\xi))\,. \tag{4.1}$$

The parameters δ^p, δ^v, and γ are patch-wise constant and depend (similar to PSPG and SUPG) on the local Peclet number. This bilinear stabilization is proposed in [3] and analyzed in [5] for the Oseen system. In particular, it was shown that for the continuous solution $u(q)$ and the discrete solution $u_h(q)$ for a fixed but arbitrary $q \in Q$ the a priori estimate (3.10) holds, if the stabilization parameters are chosen appropriately.

The following lemma states the fact that *optimize–discretize* and *discretize–optimize* coincide for the corresponding scheme:

COROLLARY 4.1. *For the LPS stabilization discretization and optimization commutes and the a priori estimate in Theorem* 3.6 *holds.*

Proof. We show firstly that the stabilization is symmetric, i.e., (P1):

$$s^{lps}(\varphi, z) = s^{lps}(z, \varphi)\,.$$

This is an immediate result of the definition of the orthogonal projection $\bar{\pi}_h$. Let us consider, firstly, the pressure stabilization ($\delta^v = \gamma = 0$). Due to $(\pi_h \nabla \xi_1, \bar{\varkappa}_h \nabla \xi_2) = 0$ for $\xi_1, \xi_2 \in Q_h$ it holds:

$$\begin{aligned} s^{lps}(\varphi, z) &= (\nabla \xi, \delta^p \bar{\varkappa}_h \nabla z^p) \\ &= (\nabla \xi, \delta^p \bar{\varkappa}_h \nabla z^p) - (\pi_h \nabla \xi, \delta^p \bar{\varkappa}_h \nabla z^p) \\ &= (\bar{\varkappa}_h \nabla \xi, \delta^p \bar{\varkappa}_h \nabla z^p) \\ &= (\bar{\varkappa}_h \nabla \xi, \delta^p \nabla z^p) \ = \ (\nabla z^p, \delta^p \bar{\varkappa}_h \nabla \xi) \\ &= s^{lps}(z, \varphi) \,. \end{aligned}$$

For $\delta^v, \gamma > 0$ the argument is the same. □

The corresponding triple norm is defined by

$$|\!|\!| u |\!|\!|_{lps} := (a(u,u) + s_h^{lps}(u,u))^{1/2},$$

fulfilling obviously the coercivity property (P2) as well as (P3) for $\sigma > 0$.

In [5] it is shown that the estimate (3.10) holds with $s = r + 1/2$ for P1, P2, Q1, and Q2 (equal-order) finite elements. Hence, we obtain the following a priori estimate.

Corollary 4.2. *For the local projection scheme in the optimal control context we obtain the following a priori estimate for* $\sigma > 0$*:*

$$|\!|\!| u - u_h |\!|\!|_{lps} \lesssim h^{r+\frac{1}{2}} (\|u\|_{r+1} + \|z\|_{r+1} + \|q\|_{r+1}) \,,$$

presumed the regularity $\{u, z, q\} \in [H^{r+1}(\Omega)]^{3d+2}$ *for the continuous solution of the optimal control problem* (2.3).

Remarks. 1. The right-hand side of the estimate involves derivatives of order $r + 1$ of the primal state, dual state, and of the control. For convection dominated flow this seems a little bit optimistic. However, the smoothness assumption above is to show that the discretization allows for a quasi-optimal convergence rate which is characteristic for stabilized methods. Even for the pure primal problem without any control it is not possible to require less regularity for obtaining a convergence order of $h^{r+1/2}$. We refer to [5] for a relaxation of the smoothness on the pressure, i.e., $p \in H^1(\Omega)$ or even less.

2. Furthermore, it should be emphasized that the whole analysis carries over to further variants of (Guermond-type) local projections, for instance,

$$s_h^{lps}(u, \varphi) := (\beta \cdot \nabla \varkappa_h v, \delta^v (\beta \cdot \nabla) \varkappa_h \phi) + (\nabla \varkappa_h p, \delta^p \nabla \varkappa_h \xi) + (\operatorname{div} \varkappa_h v, \gamma (\operatorname{div} \varkappa_h \xi)) \,,$$

where $\varkappa_h = i - i_{2h}$ is the difference between the identity and the nodal interpolation i_{2h} on the mesh $\mathcal{T}_{2h}$.

4.2. Edge stabilization. The interior penalty method of Douglas and Dupont (see [13]) was extended by Burman, Fernandez, and Hansbo to the Oseen system [10]. The principal idea is to add least squares penalty terms on the gradient jumps of velocity and pressure between neighboring elements to the Galerkin formulation. A jump term on an interior edge $E \subset \partial K$, $E \cap \partial\Omega = \emptyset$, of a cell $K \in \mathcal{T}_h$ with opposite neighbor K' is defined by

$$[\![u(x)]\!] := u(x)|_K - u(x)|_{K'} \,.$$

Jump terms on the boundary $\partial\Omega$ are set to zero. The stabilization

$$\begin{aligned} s_h^{es}(u,\varphi) &:= s_h^{es,v}(v,\phi) + s_h^{es,p}(p,\xi)\,,\\ s_h^{es,v}(v,\phi) &:= \sum_{K\in\mathcal{T}_h}\int_{\partial K}\Big\{\delta_K[\![n\cdot\nabla v]\!]\cdot[\![n\cdot\nabla\phi]\!] + \gamma_K[\![\operatorname{div} v]\!]\cdot[\![\operatorname{div}\phi]\!]\Big\}ds\,,\\ s_h^{es,p}(p,\xi) &:= \sum_{K\in\mathcal{T}_h}\int_{\partial K}\alpha_K[\![\nabla p]\!]\cdot[\![\nabla\xi]\!]\,ds\,, \end{aligned}$$

with cell dependent parameters $\alpha_K, \delta_K, \gamma_K \geq 0$ is obviously symmetric; hence, (P1) holds. The Galerkin bilinear form deviates somewhat from (3.1) due to weak implementation of boundary conditions by the Nitsche method [18]. We denote the corresponding bilinear form by $a^{es}(\cdot,\cdot)$ and refer to [10] for details. However, this does not harm the applicability of the framework in this paper to this edge stabilization technique.

It was shown in [10] that the system is stable and quasi-optimal (3.10), in the semi-norm

$$|\!|\!|u|\!|\!|'_{es} := \left(\sigma\|v\|^2 + \nu\|\nabla v\|^2 + s_h^{es,v}(v,v) + |\!|\!|v|\!|\!|^2_{\partial\Omega}\right)^{1/2}\,,$$

with the boundary contributions

$$|\!|\!|v|\!|\!|_{\partial\Omega} := \Big(\||\beta\cdot n|^{1/2}v\|^2_{\partial\Omega} + \|(\gamma\nu)^{1/2}h^{-1/2}v\|^2_{\partial\Omega} + \|\nu^{3/2}h^{-1/2}\max(Pe^{1/2},1)v\cdot n\|^2_{\partial\Omega}\Big)^{1/2}$$

and the local Peclet number $Pe = \nu^{-1}h|\beta|$. Furthermore, the stabilized system is quasi-optimal (for given control q) in the following sense:

$$|\!|\!|u(q) - u_h(q)|\!|\!|'_{es} \lesssim h^{r+1/2}\|u\|_{r+1}\,.$$

However, in order to apply Theorem 3.6 we have to consider a triple norm which allows for a coercivity property of the form (3.8). This is not valid for $|\!|\!|\cdot|\!|\!|'_{es}$, but for a triple norm including jumps of velocity gradients and pressure gradients as well:

$$|\!|\!|u|\!|\!|_{es} := \left(\sigma\|v\|^2 + \nu\|\nabla v\|^2 + |\!|\!|v|\!|\!|^2_{\partial\Omega} + s_h^{es}(u,u)\right)^{1/2}\,.$$

In [10] it was shown:

$$|\!|\!|u_h|\!|\!|'_{es} \lesssim (a^{es}(u_h,u_h) + s_h^{es,v}(v_h,v_h))^{1/2} \qquad \forall u \in X_h\,.$$

Hence, we get (P2) due to

$$\begin{aligned} |\!|\!|u_h|\!|\!|^2_{es} &= (|\!|\!|u_h|\!|\!|'_{es})^2 + s_h^{es,p}(u_h,u_h)\\ &\lesssim a^{es}(u_h,u_h) + s_h^{es,v}(v_h,v_h) + s_h^{es,p}(p_h,p_h)\\ &= a^{es}(u_h,u_h) + s_h^{es}(u_h,u_h)\,. \end{aligned}$$

Property (P3) is obviously fulfilled for $\sigma > 0$. Finally, we have to verify property (P4), i.e., the estimate (3.10) for the triple norm $|\!|\!|u|\!|\!|_{es}$:

$$|\!|\!|u(q) - u_h(q)|\!|\!|_{es} \lesssim h^{r+1/2}\|u\|_{r+1}\,.$$

But this is a direct consequence of

$$|\!|\!|u(q) - u_h(q)|\!|\!|'_{es} \lesssim h^{r+1/2}\|u\|_{r+1}\,, \tag{4.2}$$

the identity $s_h^{es,p}(p-p_h,p-p_h)=s_h^{es,p}(p_h,p_h)$, for $p=p(q)$, $p_h=p_h(q)$, and

$$s_h^{es,p}(p_h,p_h)\lesssim h^{2r+1}\|u\|_{r+1}^2\,. \tag{4.3}$$

For the exact dependence of the constant in terms of the parameters σ, β, and ν in the estimates (4.2) and (4.3) we refer to [10].

In summary, we obtain the following result.

COROLLARY 4.3. *For the edge stabilization discretization and optimization commutes and the a priori estimate*

$$|\!|\!| u-u_h |\!|\!|_{es}\lesssim h^{r+\frac{1}{2}}(\|u\|_{r+1}+\|z\|_{r+1}+\|q\|_{r+1})$$

holds if $\sigma>0$, presuming the regularity $\{u,z,q\}\in[H^{r+1}(\Omega)]^{3d+2}$ for the continuous solution of the optimal control problem (2.3).

5. Extension to Navier–Stokes.

The proposed stabilizations in sections 4.1 and 4.2 can be extended to the Navier–Stokes system. The corresponding semilinear form is

$$a(u)(\varphi):=(\operatorname{div}v,\xi)+(\sigma v,\phi)+(v\cdot\nabla v,\phi)+(\mu\nabla v,\nabla\phi)--(p,\operatorname{div}\phi)\,,$$

and the local projection stabilization term which works fine in practice, see [7], is

$$s_h^{lps}(u,\varphi):=(v\cdot\nabla v,\delta^v\bar{\varkappa}_h(v\cdot\nabla)\phi)+(\nabla p,\delta^p\bar{\varkappa}_h\nabla\xi)+(\operatorname{div}v,\gamma\bar{\varkappa}_h(\operatorname{div}\xi))\,.$$

Although these terms remain symmetric, the stabilization does not remain linear due to the nonlinearity in the convective term. Let us take $\delta^p=\gamma=0$ and assume δ^v to be independent of v in order to concentrate on the problematic term:

$$\begin{aligned}\partial_u s_h^{lps}(u)(\varphi,z)&=(v\cdot\nabla\phi,\delta^v\bar{\varkappa}_h(v\cdot\nabla)z^v)+(\phi\cdot\nabla v,\delta^v\bar{\varkappa}_h(v\cdot\nabla)z^v)\\&\quad+(v\cdot\nabla v,\delta^v\bar{\varkappa}_h(\phi\cdot\nabla)z^v)\,.\end{aligned}$$

The first term on the right-hand side is due to the orthogonality property of $\bar{\pi}_h$ equal to

$$s_h^{lps}(z,\varphi)=(v\cdot\nabla z^v,\delta^v\bar{\varkappa}_h(v\cdot\nabla)\phi)\;=\;(v\cdot\nabla\phi,\delta^v\bar{\varkappa}_h(v\cdot\nabla)z^v)\,.$$

Due to the other two remaining terms, *optimization* and *discretization* do not longer commutate. However, one may use the full gradient instead of the streamline derivative. In this case, the projection of the divergence is no longer necessary, so that we end up with the following bilinear stabilization, cf. [5]:

$$s_h^{lps}(u,\varphi):=(\nabla v,\delta^v\bar{\varkappa}_h\nabla\phi)+(\nabla p,\delta^p\bar{\varkappa}_h\nabla\xi)\,.$$

If the parameters δ^v and δ^p are independent of u (or u_h), Lemma 3.1 remains valid. However, the analysis of the correct calibration of the constants suggests a cell-wise v-dependency. In particular, if $u\in H^2(\Omega)$, one may take $\delta^v=0$ in the case of low local Reynolds number, and $\delta^v\sim h_K(1+\|v\|_{K,\infty}+h_K\sigma)$ for the convection dominant case. In the latter case, there remains a certain nonlinearity in the definition of $s_h^{lps}(u,\varphi)$, so that a minor inconsistency remains for the optimal control problem. However, since the nonlinearity is only in the stabilization parameter, it can be expected that this inconsistency is by far less problematic than residual-based stabilization.

6. Summary. In this work, we discuss the difference of *optimize-discretize* and *discretize-optimize* for optimal control problems of flow problems. In particular, we investigate the impact of stabilized finite element discretizations. We set up some sufficient conditions for stabilized finite elements for the Oseen system which lead to a consistent optimal control problem in the sense that discretization and optimization commutes. For those schemes we prove an a priori estimate. Finally, we gave two examples of stabilization, namely, local projection and edge stabilization, fitting into this concept.

REFERENCES

[1] F. Abraham, M. Behr, and M. Heinkenschloss, *The effect of stabilization in finite element methods for the optimal boundary control of the Oseen equations,* Finite Elem. Anal. Des., 41 (2004), pp. 229–251.

[2] R. Becker and M. Braack, *A finite element pressure gradient stabilization for the Stokes equations based on local projections,* Calcolo, 38 (2001), pp. 173–199.

[3] R. Becker and M. Braack, *A two-level stabilization scheme for the Navier-Stokes equations,* in Numer. Math. Adv. Appl., E. A. M. Feistauer, ed., ENUMATH 2003, Springer, 2004, pp. 123–130.

[4] R. Becker and B. Vexler, *Optimal control of the convection-diffusion equation using stabilized finite element methods,* Numer. Math., 106 (2007), pp. 349–367.

[5] M. Braack and E. Burman, *Local projection stabilization for the Oseen problem and its interpretation as a variational multiscale method,* SIAM J. Numer. Anal., 43 (2006), pp. 2544–2566.

[6] M. Braack, E. Burman, V. John, and G. Lube, *Stabilized finite element methods for the generalized Oseen problem,* Comput. Methods Appl. Mech. Engrg., 196 (2007), pp. 853–866.

[7] M. Braack and T. Richter, *Solutions of* 3*D Navier-Stokes benchmark problems with adaptive finite elements,* Computers and Fluids, 35 (2006), pp. 372–392.

[8] M. Braack and T. Richter, *Stabilized finite elements for* 3*D reactive flow,* Int. J. Numer. Methods Fluids, 51 (2006), pp. 981–999.

[9] A. Brooks and T. Hughes, *Streamline upwind Petrov-Galerkin formulation for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations,* Comput. Methods Appl. Mech. Engrg., 32 (1982), pp. 199–259.

[10] E. Burman, M. Fernandez, and P. Hansbo, *Edge stabilization for the incompressible Navier–Stokes equations: A continuous interior penalty finite element method,* SIAM J. Numer. Anal., 44 (2006), pp. 1248–1274.

[11] S. Collis and M. Heinkenschloss, *Analysis of the streamline upwind/Petrov Galerkin method applied to the solution of optimal control problems,* Technical report 02-01, Rice University, Houston, TX, 2002.

[12] L. Dedé and Q. Quarteroni, *Optimal control and numerical adaptivity for advection-diffusion equations,* Modél. Math. Anal. Numér., 39 (2005), pp. 1019–1040.

[13] J. Douglas and T. Dupont, *Interior penalty procedures for elliptic and parabolic Galerkin methods,* in Computing Methods in Applied Sciences (2nd Int. Symp., Versailles, 1975), Lecture Notes in Physics 58, Springer, New York, 1976, pp. 207–216.

[14] T. Hughes, L. Franca, and M. Balestra, *A new finite element formulation for computational fluid dynamics: V. circumvent the Babuska-Brezzi condition: A stable Petrov-Galerkin formulation for the Stokes problem accommodating equal order interpolation,* Comput. Methods Appl. Mech. Engrg., 59 (1986) pp. 89–99.

[15] C. Johnson and J. Saranen, *Streamline diffusion methods for the incompressible Euler and Navier-Stokes equations,* Math. Comput., 47 (1986), pp. 1–18.

[16] S. Li and L. Petzold, *Adjoint sensitivity analysis for time-dependent partial differential equations with adaptive mesh refinement,* J. Comput. Phys., 198 (2004), pp. 310–325.

[17] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations,* Number 170 in Die Grundlehren der mathematischen Wissenschaften, Springer-Verlag, Berlin, 1971.

[18] J. Nitsche, *Über ein Variationsprinzip zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind,* Abh. Math. Sem. Univ. Hamburg, 36 (1970/71), pp. 9–15.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 2, pp. 688–718

© 2009 Society for Industrial and Applied Mathematics

FIRST- AND SECOND-ORDER OPTIMALITY CONDITIONS FOR A CLASS OF OPTIMAL CONTROL PROBLEMS WITH QUASILINEAR ELLIPTIC EQUATIONS*

EDUARDO CASAS† AND FREDI TRÖLTZSCH‡

Abstract. A class of optimal control problems for quasilinear elliptic equations is considered, where the coefficients of the elliptic differential operator depend on the state function. First- and second-order optimality conditions are discussed for an associated control-constrained optimal control problem. Main emphasis is laid on second-order sufficient optimality conditions. To this aim, the regularity of the solutions to the state equation and its linearization is studied in detail and the Pontryagin maximum principle is derived. One of the main difficulties is the nonmonotone character of the state equation.

Key words. optimal control, distributed control, quasilinear elliptic equation, Pontryagin maximum principle, second-order optimality conditions

AMS subject classifications. 49J20, 35J65

DOI. 10.1137/080720048

1. Introduction.

In this paper, we consider optimal control problems for a quasilinear elliptic equation of the type

$$\begin{cases} -\mathrm{div}\,[a(x,y(x))\,\nabla y(x)] + f(x,y(x)) = u(x) & \text{in} \quad \Omega, \\ y(x) = 0 & \text{on} \quad \Gamma. \end{cases} \tag{1.1}$$

Equations of this type occur, for instance, in models of heat conduction, where the heat conductivity a depends on the spatial coordinate x and on the temperature y. For instance, the heat conductivity of carbon steel depends on the temperature and also on the alloying additions contained; cf. Bejan [2]. If the different alloys of steel are distributed smoothly in the domain, then $a = a(x,y)$ should depend in a sufficiently smooth way on (x,y). Similarly, the heat conductivity depends on (x,y) in the growth of silicon carbide bulk single crystals; see Klein et al. [22].

If a is independent of x, then the well-known Kirchhoff transformation is helpful to solve (1.1) uniquely. Also in the more general case $a = a(x,y)$, a Kirchhoff-type transformation can be applied. Here, we may define $b(x,y) := \int_0^y a(x,z)dz$ and set $\theta(x) := b(x,y(x))$. Under this transformation, we obtain a semilinear equation of the type $-\Delta\,\theta + \mathrm{div}\,[(\nabla_x b)(x,b^{-1}(x,\theta))] + f(x,b^{-1}(x,\theta)) = u$. We thank an anonymous referee for this hint. However, b should at least be Lipschitz with respect to x and, due to the new divergence term, the analysis of this equation is certainly not easy, too. We believe that the direct discussion of the quasilinear equation is not more difficult. Moreover, the form (1.1) seems to be more directly accessible to a numerical solution.

*Received by the editors April 2, 2008; accepted for publication (in revised form) October 31, 2008; published electronically February 13, 2009.
http://www.siam.org/journals/sicon/48-2/72004.html

†Dpto. de Matemática Aplicada y Ciencias de la Computación, E.T.S.I. Industriales y de Telecomunicación, Universidad de Cantabria, 39005 Santander, Spain (eduardo.casas@unican.es). This author was supported by the Spanish Ministry of Education and Science under projects MTM2005-06817 and "Ingenio Mathematica (i-MATH)" CSD2006-00032 (Consolider Ingenio 2010).

‡Institut für Mathematik, Technische Universität Berlin, D-10623 Berlin, Germany (troeltzsch@math.tu-berlin.de). This author was supported by the DFG research center "Mathematics for Key Technologies" (Matheon) in Berlin.

In the case $a = a(x, y)$, in spite of the nonmonotone character of the equation (1.1), there exists a celebrated comparison principle proved by Douglas, Dupont, and Serrin [16] that leads to the uniqueness of a solution of (1.1); for a more recent paper, extending this result the reader is referred to Křížek and Liu [23]. We will use the approach of [23] to deduce that (1.1) is well posed under less restrictive assumptions than those considered by the previous authors.

For other classes of quasilinear equations, in particular for equations in which a depends on the gradient of y, we refer the reader to, for instance, Lions [24] and Nečas [27].

As far as optimization is concerned, there exists a rich literature on the optimal control of semilinear elliptic and parabolic equations. For instance, the Pontryagin principle was discussed for different elliptic problems in [5], [4], [1], while the parabolic case was investigated in [6] and [29]. Problems with quasilinear equations with nonlinearity of gradient type were considered by [7], [8], [11], and [12]. This list on first-order necessary optimality conditions is by far not exhaustive. However, to our knowledge, the difficult issue of second-order conditions for problems with quasilinear equations has not yet been studied.

There is some recent progress in the case of semilinear equations. Quite a number of contributions to second-order necessary and/or sufficient optimality conditions were published for problems with such equations. We mention only [3], [14], and the state-constrained case in [10], [15], [28].

Surprisingly, the important state equation (1.1) has not yet been investigated in the context of optimal control. Our paper is the first step towards a corresponding numerical analysis. We are convinced that our analysis can also be extended to other quasilinear equations or associated systems, since the main difficulties are already inherent in (1.1).

First-order optimality conditions are needed to deduce regularity properties of optimal controls as an important prerequisite for all further investigations. The second-order analysis is a key tool for the numerical analysis of nonlinear optimal control problems. As in the minimization of a function $f : \mathbb{R} \to \mathbb{R}$, second-order sufficient conditions are commonly assumed to guarantee stability of locally optimal controls with respect to perturbations of the problem. For instance, an approximation of the PDEs by finite elements is a typical perturbation of a control problem. Associated error estimates for local solutions of the FEM-approximated optimal control problem are based on second-order sufficiency. Likewise, the standard assumption for the convergence of higher order numerical optimization algorithms such as SQP-type methods is a second-order sufficient condition at the local solution to which the method should converge.

A review on important applications of optimal control theory to problems in engineering and medical science shows that in most of the cases the underlying PDEs are quasilinear. Although our equation has a particular type, our problem might serve as a model case for the numerical analysis of optimal control problems with more general quasilinear equations or systems.

The theory of optimality conditions of associated control problems is the main issue of our paper, which is organized as follows:

First, we discuss the well-posedness of this equation in different spaces. Next, the differentiability properties of the control-to-state mapping are investigated. Based on these results, the Pontryagin maximum principle is derived. Moreover, second-order necessary and sufficient optimality conditions are established.

Notation. By $B_X(x, r)$ we denote the open ball in a normed space X with radius r centered at x, and by $\bar{B}_X(x, r)$ we denote its closure. In some formulas, the partial derivative $\partial/\partial x_j$ is sometimes abbreviated by ∂_j. By c (without index), generic constants are denoted. Moreover, $\langle \cdot\,, \cdot\rangle$ stands for the pairing between $H_0^1(\Omega)$ and $H^{-1}(\Omega)$.

2. Study of the quasilinear equation.

2.1. Existence, uniqueness, and regularity of solutions. The proof of the existence and uniqueness of a solution of (1.1) relies on the following assumptions:

(A1) $\Omega \subset \mathbb{R}^n$ is an open bounded set with a Lipschitz boundary Γ.

(A2) The functions $a : \Omega \times \mathbb{R} \to \mathbb{R}$ and $f : \Omega \times \mathbb{R} \to \mathbb{R}$ are Carathéodory, f is monotone nondecreasing with respect to the second variable for almost all $x \in \Omega$, and

$$\exists \alpha_0 > 0 \text{ such that } a(x, y) \geq \alpha_0 \text{ for a.e. } x \in \Omega \text{ and } \forall\, y \in \mathbb{R}. \tag{2.1}$$

The function $a(\cdot, 0)$ belongs to $L^\infty(\Omega)$, and for any $M > 0$ there exist a constant $C_M > 0$ and a function $\phi_M \in L^q(\Omega)$, with $q \geq pn/(n + p)$ and $n < p$, such that for all $|y|, |y_i| \leq M$

$$\begin{aligned} |a(x, y_2) - a(x, y_1)| &\leq C_M |y_2 - y_1| \text{ and} \\ |f(x, y)| &\leq \phi_M(x) \text{ for a.e. } x \in \Omega. \end{aligned} \tag{2.2}$$

In the rest of the paper q and $p \in (n, +\infty)$ will be fixed. Let us remark that $q \geq pn/(n + p) > n/2$.

Example 2.1. The following equation satisfies our assumptions if we assume $\phi_0, \phi_1 \in L^\infty(\Omega)$, $\phi_0(x) \geq \alpha_0 > 0$ a.e. in Ω, $\phi_1(x) \geq 0$ a.e. in Ω, and $1 \leq m \in \mathbb{N}$:

$$\begin{cases} -\text{div}\left[(\phi_0(x) + y^{2m}(x))\, \nabla y(x)\right] + \phi_1(x) \exp(y(x)) = u(x) & \text{in } \Omega, \\ \hfill y(x) = 0 & \text{on } \Gamma. \end{cases}$$

Theorem 2.2. *Under the assumptions* (A1) *and* (A2)*, for any element* $u \in W^{-1,p}(\Omega)$ *problem* (1.1) *has a unique solution* $y_u \in H_0^1(\Omega) \cap L^\infty(\Omega)$*. Moreover there exists* $\mu \in (0, 1)$ *independent of* u *such that* $y_u \in C^\mu(\bar{\Omega})$ *and for any bounded set* $U \subset W^{-1,p}(\Omega)$

$$\|y_u\|_{H_0^1(\Omega)} + \|y_u\|_{C^\mu(\bar{\Omega})} \leq C_U \quad \forall u \in U \tag{2.3}$$

for some constant $C_U > 0$*. Finally, if* $u_k \to u$ *in* $W^{-1,p}(\Omega)$*, then* $y_{u_k} \to y_u$ *in* $H_0^1(\Omega) \cap C^\mu(\bar{\Omega})$.

Proof. Existence of a solution. Depending on $M > 0$, we introduce the truncated function a_M by

$$a_M(x, y) = \begin{cases} a(x, y), & |y| \leq M, \\ a(x, +M), & y > +M, \\ a(x, -M), & y < -M. \end{cases}$$

In the same way, we define the truncation f_M of f. Let us prove that the equation

$$\begin{cases} -\text{div}\left[a_M(x, y)\, \nabla y\right] + f_M(x, y) = u \text{ in } \Omega, \\ \hfill y = 0 \text{ on } \Gamma \end{cases} \tag{2.4}$$

admits at least one solution $y \in H_0^1(\Omega)$. We define, for fixed $u \in W^{-1,p}(\Omega)$ and $M > 0$, a mapping $F : L^2(\Omega) \to L^2(\Omega)$ by $F(z) = y$, where $y \in H_0^1(\Omega)$ is the unique solution to

$$\begin{cases} -\mathrm{div}\,[a_M(x,z)\,\nabla y] + f_M(x,z) = u \text{ in } \Omega, \\ \hfill y = 0 \text{ on } \Gamma. \end{cases} \tag{2.5}$$

Thanks to assumption (A2), (2.2), we have

$$|f_M(x,z)| \leq \phi_M(x)$$

and $\phi_M \in L^q(\Omega) \subset H^{-1}(\Omega)$. Therefore, (2.5) is a linear equation and $u - f_M(\cdot, z)$ belongs to $H^{-1}(\Omega)$; hence (2.5) admits a unique solution $y_M \in H_0^1(\Omega)$ and F is well defined. It can be shown by standard arguments invoking in particular the compact injection of $H^1(\Omega)$ in $L^2(\Omega)$ that F is continuous. Furthermore, we have

$$\|y_M\|_{H^1(\Omega)} \leq \frac{1}{\alpha_0} \left(\|u\|_{H^{-1}(\Omega)} + \|\phi_M\|_{H^{-1}(\Omega)}\right). \tag{2.6}$$

Using this estimate and the fact that $H^1(\Omega)$ is compactly embedded into $L^2(\Omega)$, it is easy to apply the Schauder theorem to prove the existence of a fixed point $y_M \in H_0^1(\Omega)$ of F. Obviously, y_M is a solution of (2.4).

Since $q \geq np/(n+p)$ we have that $L^q(\Omega) \subset W^{-1,p}(\Omega)$. Now an application of the Stampacchia truncation method yields

$$\|y_M\|_{L^\infty(\Omega)} \leq c_\infty \, \|u - f(\cdot, 0)\|_{W^{-1,p}(\Omega)}, \tag{2.7}$$

where c_∞ depends only on the coercivity constant α_0 given in (2.1) but neither on $\|a_M(\cdot, y_M)\|_{L^\infty(\Omega)}$ nor on $f_M(\cdot, y_M)$. For the idea of this method, the reader is referred to Stampacchia [30] or to the exposition for semilinear elliptic equations in Tröltzsch [31, Theorem 7.3]. By taking

$$M \geq c_\infty \, \|u - f(\cdot, 0)\|_{W^{-1,p}(\Omega)},$$

(2.7) implies that $a_M(x, y_M(x)) = a(x, y_M(x))$ and $f_M(x, y_M(x)) = f(x, y_M(x))$ for a.e. $x \in \Omega$, and therefore $y_M \in H_0^1(\Omega) \cap L^\infty(\Omega)$ is a solution of (1.1). The Hölder regularity follows as usual; see, for instance, Gilbarg and Trudinger [19, Theorem 8.29]. Inequality (2.3) follows from (2.6), (2.7), and the estimates in [19, Theorem 8.29]. Finally, the convergence property can be deduced from (2.3) easily once the uniqueness is proved.

Uniqueness of a solution. Here we follow the method by Křížek and Liu [23]. Let us assume that $y_i \in H_0^1(\Omega) \cap L^\infty(\Omega)$, $i = 1, 2$, are two solutions of (1.1). The regularity results proved above imply that $y_i \in C(\bar{\Omega})$, $i = 1, 2$. Let us define the open sets

$$\Omega_0 = \{x \in \Omega : y_2(x) - y_1(x) > 0\}$$

and for every $\varepsilon > 0$

$$\Omega_\varepsilon = \{x \in \Omega : y_2(x) - y_1(x) > \varepsilon\}.$$

Now we take $z_\varepsilon(x) = \min\{\varepsilon, (y_2(x) - y_1(x))^+\}$, which belongs to $H_0^1(\Omega)$ and $|z_\varepsilon| \leq \varepsilon$. Multiplying the equations corresponding to y_i by z_ε and doing the usual integration by parts we get

$$\int_\Omega \{a(x, y_i)\nabla y_i \cdot \nabla z_\varepsilon + f(x, y_i) z_\varepsilon\}\, dx = \langle u, z_\varepsilon \rangle, \quad i = 1, 2.$$

By subtracting both equations, using the monotonicity of f, (2.1) and (2.2) and the fact that $\nabla z_\varepsilon(x) = 0$ for a.a. $x \notin \Omega_0 \backslash \Omega_\varepsilon$ and in view of $\nabla z_\varepsilon = \nabla(y_2 - y_1)^+ = \nabla(y_2 - y_1)$ a.e. in $\Omega_0 \setminus \Omega_\varepsilon$ we get

$$\alpha_0 \|\nabla z_\varepsilon\|^2_{L^2(\Omega)} \leq \int_\Omega \{a(x,y_2)|\nabla z_\varepsilon|^2 + [f(x,y_2) - f(x,y_1)]z_\varepsilon\}dx$$

$$= \int_\Omega \{a(x,y_2)\nabla(y_2 - y_1) \cdot \nabla z_\varepsilon + [f(x,y_2) - f(x,y_1)]z_\varepsilon\}dx$$

and, invoking the weak formulation of the equation for y_1,

$$= \int_\Omega [a(x,y_1)\nabla y_1 - a(x,y_2)\nabla y_1] \cdot \nabla z_\varepsilon \, dx$$

$$= \int_{\Omega_0 \backslash \Omega_\varepsilon} [a(x,y_1)\nabla y_1 - a(x,y_2)\nabla y_1] \cdot \nabla z_\varepsilon \, dx$$

$$\leq C_M \|y_2 - y_1\|_{L^\infty(\Omega_0 \backslash \Omega_\varepsilon)} \|\nabla y_1\|_{L^2(\Omega_0 \backslash \Omega_\varepsilon)} \|\nabla z_\varepsilon\|_{L^2(\Omega_0 \backslash \Omega_\varepsilon)}$$

$$\leq C_M \varepsilon \|\nabla y_1\|_{L^2(\Omega_0 \backslash \Omega_\varepsilon)} \|\nabla z_\varepsilon\|_{L^2(\Omega_0 \backslash \Omega_\varepsilon)}.$$

From this inequality, along with Friedrich's inequality, we get

$$\|z_\varepsilon\|_{L^2(\Omega)} \leq C\|\nabla z_\varepsilon\|_{L^2(\Omega)} \leq C'\varepsilon \|\nabla y_1\|_{L^2(\Omega_0 \backslash \Omega_\varepsilon)}. \tag{2.8}$$

Now by $\lim_{\varepsilon \downarrow 0} |\Omega_0 \setminus \Omega_\varepsilon| = 0$ and (2.8) we deduce

$$|\Omega_\varepsilon| = \varepsilon^{-2} \int_{\Omega_\varepsilon} \varepsilon^2 \leq \varepsilon^{-2} \int_\Omega |z_\varepsilon|^2 \leq C'' \|\nabla y_1\|^2_{L^2(\Omega_0 \backslash \Omega_\varepsilon)} \to 0,$$

which implies that $|\Omega_0| = \lim_{\varepsilon \to 0} |\Omega_\varepsilon| = 0$ and hence $y_2 \leq y_1$. In the same way, we prove that $y_1 \leq y_2$ □

As in this theorem, throughout our paper, the solutions of PDEs are defined as weak solutions.

Remark 2.3. Let us remark that the Lipschitz property of a with respect to y assumed in (A2) was necessary only to prove the uniqueness of a solution of (1.1), but it was not needed to establish existence and regularity. We can get multiple solutions of (1.1) if the Lipschitz property (2.2) fails; see Hlaváček, Křížek, and Malý [21] for a one-dimensional example.

By assuming more regularity on a, f, Γ, and u, we can obtain higher regularity of the solutions of (1.1).

THEOREM 2.4. *Let us suppose that* (A1) *and* (A2) *hold. We also assume that* $a : \bar{\Omega} \times \mathbb{R} \longrightarrow \mathbb{R}$ *is continuous and* Γ *is of class* C^1. *Then, for any* $u \in W^{-1,p}(\Omega)$, (1.1) *has a unique solution* $y_u \in W_0^{1,p}(\Omega)$. *Moreover, for any bounded set* $U \subset W^{-1,p}(\Omega)$, *there exists a constant* $C_U > 0$ *such that*

$$\|y_u\|_{W_0^{1,p}(\Omega)} \leq C_U \quad \forall u \in U. \tag{2.9}$$

If $u_k \to u$ *in* $W^{-1,p}(\Omega)$, *then* $y_{u_k} \to y_u$ *strongly in* $W_0^{1,p}(\Omega)$.

The proof of this theorem follows from Theorem 2.2 and the $W^{1,p}(\Omega)$-regularity results for linear elliptic equations; see Giaquinta [18, Chap. 4, p. 73] or Morrey [25, pp. 156–157]. It is enough to remark that the function $\hat{a}(x) = a(x, y_u(x))$ is continuous in $\bar{\Omega}$ and $u - f(\cdot, y_u)$ belongs to $W^{-1,p}(\Omega)$.

Let us state some additional assumptions leading to $W^{2,q}(\Omega)$-regularity for the solutions of (1.1).

(A3) For all $M > 0$, there exists a constant $c_M > 0$ such that the following local Lipschitz property is satisfied:

$$|a(x_1, y_1) - a(x_2, y_2)| \leq c_M \ \{|x_1 - x_2| + |y_1 - y_2|\} \tag{2.10}$$

for all $x_i \in \bar{\Omega}$, $y_i \in [-M, M]$, $i = 1, 2$.

THEOREM 2.5. *Under the hypotheses* (A1)–(A3) *and assuming that* Γ *is of class* $C^{1,1}$, *for any* $u \in L^q(\Omega)$, (1.1) *has one solution* $y_u \in W^{2,q}(\Omega)$. *Moreover, for any bounded set* $U \subset L^q(\Omega)$, *there exists a constant* $C_U > 0$ *such that*

$$\|y_u\|_{W^{2,q}(\Omega)} \leq C_U \quad \forall u \in U. \tag{2.11}$$

Proof. (i) From Sobolev embedding theorems (cf. Nečas [26, Theorem 3.4]), it follows that

$$L^q(\Omega) \hookrightarrow W^{-1,\frac{nq}{n-q}}(\Omega) \ \text{ if } \ 1 < q < n, \tag{2.12}$$

$$L^q(\Omega) \hookrightarrow W^{-1,\infty}(\Omega) \ \text{ if } \ n \leq q < \infty. \tag{2.13}$$

Since $L^q(\Omega) \subset W^{-1,p}(\Omega)$, we can apply Theorem 2.4 to get the existence of at least one solution in $W_0^{1,p}(\Omega)$ for every $1 < p < \infty$ if $q \geq n$, and for $p = \frac{nq}{n-q}$ if $q < n$. We have to prove the $W^{2,q}(\Omega)$-regularity. To this aim, we distinguish between two cases in the proof.

(ii)(a) *Case* $q \geq n$. We have that $y \in W_0^{1,p}(\Omega)$ for any $p < \infty$, in particular in $W_0^{1,2q}(\Omega)$. By using assumption (A3), expanding the divergence term of the PDE (1.1), and dividing by a we find that

$$-\Delta y = \underbrace{\frac{1}{a}}_{L^\infty} \left\{ \underbrace{u - f(\cdot, y)}_{L^q} + \sum_{j=1}^{n} \underbrace{\partial_j a(x, y)}_{L^\infty} \underbrace{\partial_j y}_{L^q} + \underbrace{\frac{\partial a}{\partial y}}_{L^\infty} \underbrace{|\nabla y|^2}_{L^q} \right\}, \tag{2.14}$$

hence the right-hand side of (2.14) is in $L^q(\Omega)$. Notice that $\frac{\partial a}{\partial y} \in L^\infty$ follows from (2.10) and the boundedness of y. The $C^{1,1}$ smoothness of Γ permits us to apply a well-known result by Grisvard [20] on maximal regularity and to get $y \in W^{2,q}(\Omega)$.

(ii)(b) *Case* $n/2 < q < n$. Notice that $y \in W_0^{1,\frac{nq}{n-q}}(\Omega)$. It follows that $|\nabla y|^2 \in L^{\frac{nq}{2(n-q)}}(\Omega)$. A simple calculation confirms that

$$\frac{nq}{2(n-q)} > q, \tag{2.15}$$

since this is equivalent to $q > n/2$, a consequence of our assumption on q. Therefore, it holds that $|\nabla y|^2 \in L^q(\Omega)$ and once again the right-hand side of (2.14) belongs to $L^q(\Omega)$. We apply again the regularity results by Grisvard [20] to obtain $y \in W^{2,q}(\Omega)$. □

COROLLARY 2.6. *Suppose that the assumptions of Theorem* 2.5, *except the regularity hypothesis of* Γ, *are satisfied with* $q = 2$. *Then, if* $\Omega \subset \mathbb{R}^n$ *is an open, bounded, and convex set,* $n = 2$ *or* $n = 3$, *there exists one solution of* (1.1): $y \in H^2(\Omega) \cap H_0^1(\Omega)$.

Proof. This is a simple extension of Theorem 2.5 for $q = 2$. Notice that we have assumed $n \leq 3$ so that $q > n/2$ is true. The $C^{1,1}$ smoothness of Γ is not needed for convex domains, since maximal regularity holds there; cf. [20]. □

2.2. Differentiability of the control-to-state mapping. In order to derive the first- and second-order optimality conditions for the control problem, we need to assume some differentiability of the functions involved in the control problem. In this section, we will analyze the differentiability properties of the states with respect to the control. To this aim, we require the following assumption.

(A4) The functions a and f are of class C^2 with respect to the second variable and, for any number $M > 0$, there exists a constant $D_M > 0$ such that

$$\sum_{j=1}^{2} \left|\frac{\partial^j a}{\partial y^j}(x,y)\right| + \left|\frac{\partial^j f}{\partial y^j}(x,y)\right| \leq D_M \quad \text{for a.e. } x \in \Omega \text{ and } \forall\ |y| \leq M.$$

Now we are going to study the differentiability of the control-to-state mapping. As a first step we study the linearized equation of (1.1) around a solution y_u. The reader should note that the well-posedness of the linearized equation is not obvious because of the linear operator is not monotone.

THEOREM 2.7. *Given $y \in W^{1,p}(\Omega)$ for any $v \in H^{-1}(\Omega)$ the linearized equation*

$$(2.16)\qquad \begin{cases} -\mathrm{div}\left[a(x,y)\nabla z + \dfrac{\partial a}{\partial y}(x,y)z\,\nabla y\right] + \dfrac{\partial f}{\partial y}(x,y)\,z = v & \text{in } \Omega, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad z = 0 & \text{on } \Gamma \end{cases}$$

has a unique solution $z_v \in H_0^1(\Omega)$.

Remark 2.8. As a consequence of the open mapping theorem, assuming that (A2) and (A4) hold, we know that the relation $v \mapsto z_v$ defined by (2.16) is an isomorphism between $H^{-1}(\Omega)$ and $H_0^1(\Omega)$. Indeed, it is enough to note that the linear mapping

$$z \mapsto -\mathrm{div}\left[a(x,y)\nabla z + \frac{\partial a}{\partial y}(x,y)z\,\nabla y\right] + \frac{\partial f}{\partial y}(x,y)\,z$$

is continuous from $H_0^1(\Omega)$ to $H^{-1}(\Omega)$. To verify this, we notice first that $a(x,y)$, $\frac{\partial a}{\partial y}(x,y)$, and $\frac{\partial f}{\partial y}(x,y)$ are bounded functions because of our assumptions and the boundedness of y, which follows from the fact that $y \in W_0^{1,p}(\Omega) \subset C(\bar{\Omega})$ for $p > n$. The only delicate point is to check that

$$\frac{\partial a}{\partial y}(\cdot,y)z\nabla y \in L^2(\Omega)^n.$$

This property follows from the Hölder inequality

$$\left(\int_\Omega \left|\frac{\partial a}{\partial y}(\cdot,y)z\nabla y\right|^2 dx\right)^{1/2} \leq D_M\|z\|_{L^{\frac{2p}{p-2}}(\Omega)}\|\nabla y\|_{L^p(\Omega)}$$

and the fact that

$$H_0^1(\Omega) \subset L^{\frac{2n}{n-2}}(\Omega) \subset L^{\frac{2p}{p-2}}(\Omega) \text{ if } n > 2,$$

$$H_0^1(\Omega) \subset L^r(\Omega)\ \forall\ r < \infty \text{ if } n = 2,$$

where we have used that

$$p > n \Rightarrow \frac{2n}{n-2} > \frac{2p}{p-2}.$$

Remark 2.9. The reader can easily check that the proof of Theorem 2.7 can be modified in a very obvious way to state that the equation

$$\begin{cases} -\mathrm{div}\left[a(x,y_1)\nabla z + \dfrac{\partial a}{\partial y}(x,y_2)z\,\nabla y\right] + \dfrac{\partial f}{\partial y}(x,y_3)\,z = v & \text{in } \Omega,\\ \hfill z = 0 & \text{on } \Gamma \end{cases}$$

has a unique solution in $z \in H_0^1(\Omega)$ for any elements $y \in W^{1,p}(\Omega)$ and $y_i \in L^\infty(\Omega)$, $i = 1, 2, 3$.

Proof of Theorem 2.7. First we prove the uniqueness and then the existence.

Uniqueness of solution of (2.16). We follow the same approach used to prove the uniqueness of a solution of (1.1). Let us take $v = 0$ and assume that $z \in H_0^1(\Omega)$ is a solution of (2.16); then the goal is to prove that $z = 0$. Thus we define the sets

$$\Omega_0 = \{x \in \Omega : z(x) > 0\} \text{ and } \Omega_\varepsilon = \{x \in \Omega : z(x) > \varepsilon\}.$$

Now we set $z_\varepsilon(x) = \min\{\varepsilon, z^+(x)\}$, so that $z_\varepsilon \in H_0^1(\Omega)$, $|z_\varepsilon| \le \varepsilon$, $zz_\varepsilon \ge 0$, $z\nabla z_\varepsilon = z_\varepsilon \nabla z_\varepsilon$, and $\nabla z \cdot \nabla z_\varepsilon = |\nabla z_\varepsilon|^2$. Then multiplying the equation corresponding to z by z_ε and performing an integration by parts we get

$$\int_\Omega \left\{a(x,y)|\nabla z_\varepsilon|^2 + \frac{\partial a}{\partial y}(x,y)z_\varepsilon \nabla y \cdot \nabla z_\varepsilon + \frac{\partial f}{\partial y}(x,y)z_\varepsilon^2\right\} dx = 0;$$

then, by the monotonicity of f and (A2),

$$\begin{aligned} \alpha_0 \|\nabla z_\varepsilon\|^2_{L^2(\Omega)} &\le \int_\Omega \left\{a(x,y)|\nabla z_\varepsilon|^2 + \frac{\partial f}{\partial y}(x,y)z_\varepsilon^2\right\} dx \\ = -\int_\Omega \frac{\partial a}{\partial y}(x,y)z_\varepsilon \nabla y \cdot \nabla z_\varepsilon \, dx &= -\int_{\Omega_0 \setminus \Omega_\varepsilon} \frac{\partial a}{\partial y}(x,y)z_\varepsilon \nabla y \cdot \nabla z_\varepsilon \, dx \\ &\le C_M \|\nabla y\|_{L^p(\Omega_0\setminus\Omega_\varepsilon)} \|\nabla z_\varepsilon\|_{L^2(\Omega)}. \end{aligned}$$

From here follows an inequality analogous to (2.8), and continuing the proof in a similar manner, we conclude that $|\Omega_0| = \lim_{\varepsilon\to 0} |\Omega_\varepsilon| = 0$, and therefore $z \le 0$ in Ω. But $-z$ is also a solution of (2.16), so by the same arguments we deduce that $-z \le 0$ in Ω, and therefore $z = 0$.

Existence of solution of (2.16). For every $t \in [0, 1]$ let us consider the equation

$$\begin{cases} -\mathrm{div}\left[a(x,y)\nabla z + t\dfrac{\partial a}{\partial y}(x,y)z\,\nabla y_u\right] + \dfrac{\partial f}{\partial y}(x,y)\,z = v & \text{in } \Omega,\\ \hfill z = 0 & \text{on } \Gamma. \end{cases} \tag{2.17}$$

For $t = 0$, the resulting linear equation is monotone, and by an obvious application of the Lax–Milgram theorem we know that there exists a unique solution $z_0 \in H_0^1(\Omega)$ for every $v \in H^{-1}(\Omega)$. Let us denote by S the set of points $t \in [0, 1]$ for which (2.17) defines an isomorphism between $H_0^1(\Omega)$ and $H^{-1}(\Omega)$. S is not empty because $0 \in S$. Let us denote by t_{max} the supremum of S. We will prove first that $t_{max} \in S$, and then we will see that $t_{max} = 1$, which concludes the proof of existence.

Let us take a sequence $\{t_k\}_{k=1}^\infty \subset S$ such that $t_k \to t_{max}$ when $k \to \infty$ and let us denote by z_k the solutions of (2.17) corresponding to the values t_k. Multiplying the equation of z_k by z_k and integrating by parts, using assumptions (A1) and (A2) we get

$$\begin{aligned}\alpha_0\|\nabla z_k\|^2_{L^2(\Omega)} &\le \int_\Omega \left\{a(x,y)|\nabla z_k|^2 + \frac{\partial f}{\partial y}(x,y)z_k^2\right\}dx\\ &= \langle v, z_k\rangle - t_k\int_\Omega \frac{\partial a}{\partial y}(x,y)z_k\nabla y\cdot\nabla z_k\,dx\\ &\le \left(\|v\|_{H^{-1}(\Omega)} + t_kD_M\|\nabla y\|_{L^p(\Omega)}\|z_k\|_{L^{\frac{2p}{p-2}}(\Omega)}\right)\|\nabla z_k\|_{L^2(\Omega)},\end{aligned}$$

which implies

$$\|\nabla z_k\|_{L^2(\Omega)} \le C\left(\|v\|_{H^{-1}(\Omega)} + \|z_k\|_{L^{\frac{2p}{p-2}}(\Omega)}\right). \tag{2.18}$$

In principle it seems that there are two possibilities: either $\{z_k\}_{k=1}^\infty$ is bounded in $L^{\frac{2p}{p-2}}(\Omega)$ or it is not. In the first case (2.18) implies that $\{z_k\}_{k=1}^\infty$ is bounded in $H_0^1(\Omega)$; then we can extract a subsequence, denoted in the same way, such that $z_k \rightharpoonup z$ weakly in $H_0^1(\Omega)$ and strongly in $L^{\frac{2p}{p-2}}(\Omega)$ because of the compactness of the embedding $H_0^1(\Omega)\subset L^{\frac{2p}{p-2}}(\Omega)$ for $p>n$. Therefore we can pass to the limit in (2.17), with $t=t_k$, and check that z is a solution of (2.17) for $t=t_{max}$, and therefore $t_{max}\in S$, as we wanted to prove.

Let us see that the second possibility is not actually a correct assumption. Indeed, let us assume that $\|z_k\|_{L^{\frac{2p}{p-2}}(\Omega)}\to\infty$, taking a subsequence if necessary. We define

$$\rho_k = \frac{1}{\|z_k\|_{L^{\frac{2p}{p-2}}(\Omega)}} \to 0 \quad\text{and}\quad \hat z_k = \rho_k z_k.$$

Then from (2.18) we deduce

$$\|\nabla \hat z_k\|_{L^2(\Omega)} \le C\left(\rho_k\|v\|_{H^{-1}(\Omega)} + \|\hat z_k\|_{L^{\frac{2p}{p-2}}(\Omega)}\right) = C\left(\rho_k\|v\|_{H^{-1}(\Omega)}+1\right). \tag{2.19}$$

Moreover $\hat z_k$ satisfies the equation

$$\begin{cases} -\text{div}\left[a(x,y)\nabla\hat z_k + t_k\dfrac{\partial a}{\partial y}(x,y)\hat z_k\,\nabla y\right] + \dfrac{\partial f}{\partial y}(x,y)\hat z_k = \rho_k v & \text{in } \Omega,\\ z = 0 & \text{on } \Gamma.\end{cases} \tag{2.20}$$

From (2.19) we know that we can extract a subsequence, denoted once again in the same way, such that $\hat z_k \rightharpoonup \hat z$ weakly in $H_0^1(\Omega)$ and strongly in $L^{\frac{2p}{p-2}}(\Omega)$. Then $\|\hat z\|_{L^{\frac{2p}{p-2}}(\Omega)} = 1$ and passing to the limit in (2.20) we have that $\hat z$ satisfies the equation

$$\begin{cases} -\text{div}\left[a(x,y)\nabla\hat z + t_{max}\dfrac{\partial a}{\partial y}(x,y)\hat z\,\nabla y\right] + \dfrac{\partial f}{\partial y}(x,y)\hat z = 0 & \text{in } \Omega,\\ z = 0 & \text{on } \Gamma.\end{cases}$$

But we have already proved the uniqueness of solution of (2.16); the fact of including t_{max} in the equation does not matter for the proof. Therefore $\hat z = 0$, which contradicts the fact that its norm in $L^{\frac{2p}{p-2}}(\Omega)$ is one.

Finally we prove that $t_{max} = 1$. If it is false, then let us consider the operators $T_\varepsilon, T_{max} \in \mathcal{L}(H_0^1(\Omega), H^{-1}(\Omega))$ for any $\varepsilon > 0$ with $t_{max} + \varepsilon \leq 1$, defined by

$$T_\varepsilon z = -\text{div}\left[a(x,y)\nabla z + (t_{max}+\varepsilon)\frac{\partial a}{\partial y}(x,y)z\,\nabla y\right] + \frac{\partial f}{\partial y}(x,y)z,$$
$$T_{max} z = -\text{div}\left[a(x,y)\nabla z + t_{max}\frac{\partial a}{\partial y}(x,y)z\,\nabla y\right] + \frac{\partial f}{\partial y}(x,y)z.$$

Then we have

$$\|T_\varepsilon - T_{max}\|_{\mathcal{L}(H_0^1(\Omega),H^{-1}(\Omega))} = \sup_{\|z\|_{H_0^1(\Omega)}\leq 1} \|(T_\varepsilon - T_{max})z\|_{H^{-1}(\Omega)}$$
$$\leq D_M \sup_{\|z\|_{H_0^1(\Omega)}\leq 1} \varepsilon\|z\|_{L^{\frac{2p}{p-2}}(\Omega)}\|\nabla y\|_{L^p(\Omega)} \leq C\varepsilon.$$

Since T_{max} is an isomorphism, if $C\varepsilon < 1$, then T_ε is also an isomorphism, which contradicts the fact that t_{max} is the supremum of S. □

THEOREM 2.10. *Let us suppose that* (A1), (A2), *and* (A4) *hold. We also assume that $a : \bar{\Omega}\times\mathbb{R} \mapsto \mathbb{R}$ is continuous and Γ is of class C^1. Then the control-to-state mapping $G : W^{-1,p}(\Omega) \to W_0^{1,p}(\Omega)$, $G(u) = y_u$, is of class C^2. Moreover, for any $v, v_1, v_2 \in W^{-1,p}(\Omega)$ the functions $z_v = G'(u)v$ and $z_{v_1,v_2} = G''(u)[v_1,v_2]$ are the unique solutions in $W_0^{1,p}(\Omega)$ of the equations*

$$\text{(2.21)} \qquad \begin{cases} -\text{div}\left[a(x,y_u)\nabla z + \dfrac{\partial a}{\partial y}(x,y_u)z\,\nabla y_u\right] + \dfrac{\partial f}{\partial y}(x,y)\,z = v & \text{in } \Omega,\\ z = 0 & \text{on } \Gamma \end{cases}$$

and

(2.22)
$$\begin{cases} -\,\text{div}\left[a(x,y_u)\nabla z + \dfrac{\partial a}{\partial y}(x,y_u)z\nabla y_u\right] + \dfrac{\partial f}{\partial y}(x,y_u)\,z = -\dfrac{\partial^2 f}{\partial y^2}(x,y_u)z_{v_1}z_{v_2}\\ \quad +\,\text{div}\left[\dfrac{\partial a}{\partial y}(x,y_u)(z_{v_1}\nabla z_{v_2} + \nabla z_{v_1}z_{v_2}) + \dfrac{\partial^2 a}{\partial y^2}(x,y_u)z_{v_1}z_{v_2}\nabla y_u\right] \quad \text{in } \Omega,\\ z = 0 \quad \text{on } \Gamma, \end{cases}$$

respectively, where $z_i = G'(u)v_i$, $i = 1, 2$.

Proof. We introduce the mapping $F : W_0^{1,p}(\Omega)\times W^{-1,p}(\Omega) \to W^{-1,p}(\Omega)$ by

$$F(y,u) = -\text{div}\,[a(\cdot,y)\nabla y] + f(\cdot,y) - u.$$

Because of the assumptions (A2) and (A4), it is obvious that F is well defined, of class C^2, and $F(y_u,u) = 0$ for every $u \in W_0^{1,p}(\Omega)$. If we prove that

$$\frac{\partial F}{\partial y}(y_u,u) : W_0^{1,p}(\Omega) \longrightarrow W^{-1,p}(\Omega)$$

is an isomorphism, then we can apply the implicit function theorem to deduce the theorem, getting (2.21) and (2.22) by simple computations. Let us remark that

$$\frac{\partial F}{\partial y}(y_u,u)z = -\text{div}\left[a(x,y_u)\nabla z + \frac{\partial a}{\partial y}(x,y_u)z\,\nabla y_u\right] + \frac{\partial f}{\partial y}(x,y_u)\,z.$$

According to Theorem 2.7, for any $v \in H^{-1}(\Omega)$, there exists a unique element $z \in H_0^1(\Omega)$ such that

$$\frac{\partial F}{\partial y}(y_u, u)z = v.$$

It is enough to prove that $z \in W_0^{1,p}(\Omega)$ if $v \in W^{-1,p}(\Omega) \subset H^{-1}(\Omega)$. More precisely, this means that the unique solution of (2.16) in $H_0^1(\Omega)$ belongs to $W_0^{1,p}(\Omega)$. First of all, let us note that

$$a(\cdot, y_u) \in L^\infty(\Omega),\ \frac{\partial a}{\partial y}(\cdot, y_u)\nabla y_u \in L^p(\Omega)^n,\ \frac{\partial f}{\partial y}(\cdot, y_u) \in L^\infty(\Omega), \quad \text{and} \quad v \in W^{-1,p}(\Omega).$$

Therefore, we can apply a result by Stampacchia [30, Theorem 4.1 and Remark 4.2] about $L^\infty(\Omega)$-estimates of solutions of linear equations to get that $z \in L^\infty(\Omega)$. Now we have that

$$-\mathrm{div}[a(x, y_u)\nabla z] = v + \mathrm{div}\left[\frac{\partial a}{\partial y}(x, y_u)z\nabla y_u\right] - \frac{\partial f}{\partial y}(x, y_u)z \in W^{-1,p}(\Omega)$$

and $x \mapsto a(x, y_u(x))$ is a continuous real-valued function defined in $\bar{\Omega}$. Finally, as in the proof of Theorem 2.4, we can use the $W_0^{1,p}(\Omega)$-regularity results for linear equations (see [18, Chap. 4, p. 73] or [25, pp. 156–157]) to deduce that $z \in W_0^{1,p}(\Omega)$. □

From Theorem 2.5 we know that the states y corresponding to controls $u \in L^q(\Omega)$, with $q > n/2$, can have an extra regularity under certain assumptions. In this situation, a natural question arises. Can we prove a result analogous to Theorem 2.10 with $G : L^q(\Omega) \to W^{2,q}(\Omega)$? The answer is positive if we assume some extra regularity of the function a.

(A5) For all $M > 0$, there exists a constant $d_M > 0$ such that the following inequality is satisfied:

$$\left|\frac{\partial^j a}{\partial y^j}(x_1, y_1) - \frac{\partial^j a}{\partial y^j}(x_2, y_2)\right| \le d_M\ \{|x_1 - x_2| + |y_1 - y_2|\} \tag{2.23}$$

for all $x_i \in \bar{\Omega}$, $y_i \in [-M, M]$, $i = 1, 2$ and $j = 1, 2$.

Theorem 2.11. *Suppose that* (A1)–(A5) *hold and* Γ *is of class* $C^{1,1}$. *Then the control-to-state mapping* $G : L^q(\Omega) \to W^{2,q}(\Omega)$, $G(u) = y_u$, *is of class* C^2. *For any* $v, v_1, v_2 \in L^q(\Omega)$, *the functions* $z_v = G'(u)v$ *and* $z_{v_1,v_2} = G''(u)[v_1, v_2]$ *are the unique solutions in* $W^{2,q}(\Omega) \cap W_0^{1,q}(\Omega)$ *of* (2.21) *and* (2.22), *respectively.*

Proof. The proof follows the same steps as in the previous theorem, with obvious modifications. Let us note the main differences. This time, the function F is defined by the same expression as above and acts from $(W^{2,q}(\Omega) \cap W_0^{1,q}(\Omega)) \times L^q(\Omega)$ to $L^q(\Omega)$. We have to check that F is well defined, and we must determine the first- and second-order derivatives. By using the assumptions (A3)–(A5), we have for $j = 0, 1, 2$ and $y \in W^{2,q}(\Omega) \cap W_0^{1,q}(\Omega)$ that

$$\begin{aligned}\mathrm{div}\left[\frac{\partial^j a}{\partial y^j}(x, y(x))\nabla y(x)\right] &= \left[\nabla_x \frac{\partial^j a}{\partial y^j}\right](x, y(x)) \cdot \nabla y(x) + \frac{\partial^{j+1} a}{\partial y^{j+1}}(x, y(x))|\nabla y(x)|^2 \\ &+ \frac{\partial^j a}{\partial y^j}(x, y(x))\Delta y(x) \in L^q(\Omega).\end{aligned} \tag{2.24}$$

We have used the fact that $(\partial^j a/\partial y^j)$ is Lipschitz in x and y, and therefore differentiable a.e., and that the chain rule is valid in the framework of Sobolev spaces.

On the other hand, (A2) and (A4) imply that

$$\frac{\partial^j f}{\partial y^j}(\cdot, y) \in L^q(\Omega) \quad \text{for} \quad j = 0, 1, 2.$$

From these remarks, it is easy to deduce that F is of class C^2. Let us prove that (2.16) has a unique solution $z \in W^{2,q}(\Omega) \cap W_0^{1,q}(\Omega)$ for any $v \in L^q(\Omega)$. The uniqueness is an immediate consequence of the uniqueness of solution in $H_0^1(\Omega) \cap L^\infty(\Omega)$. It remains to prove the $W^{2,q}$-regularity. We argue similarly to the proof of Theorem 2.4. From (2.16) we get

$$\begin{aligned} -\Delta z &= \frac{1}{a}\left\{ v + \operatorname{div}\left[\frac{\partial a}{\partial y}(x,\bar{y})\, z\, \nabla \bar{y}\right] - \frac{\partial f}{\partial y}(x,\bar{y})\, z + v \right\} + \nabla_x a \cdot \nabla z + \frac{\partial a}{\partial y} \nabla \bar{y} \cdot \nabla z \\ &= \frac{1}{a}\left\{ v - \frac{\partial f}{\partial y}(x,\bar{y})\, z + \nabla_x \frac{\partial a}{\partial y}\, z \cdot \nabla \bar{y} + \frac{\partial^2 a}{\partial y^2}\, z\, |\nabla \bar{y}|^2 + \frac{\partial a}{\partial y} \nabla z \cdot \nabla \bar{y} + \frac{\partial a}{\partial y}\, z\, \Delta \bar{y} \right\} \\ &\quad + \nabla_x a \cdot \nabla z + \frac{\partial a}{\partial y}\, \nabla z \cdot \nabla \bar{y}. \end{aligned}$$

The right-hand side is an element of $L^q(\Omega)$. To verify this, consider, for instance, the term with the lowest regularity, i.e., the term $\nabla \bar{y} \cdot \nabla z$:

$$\begin{aligned} \left(\int_\Omega |\nabla \bar{y}|^q |\nabla z|^q dx \right)^{\frac{1}{q}} &\le \left(\int_\Omega |\nabla \bar{y}|^n dx \right)^{\frac{1}{n}} \left(\int_\Omega |\nabla z|^{\frac{nq}{n-q}} dx \right)^{\frac{n-q}{nq}} \\ &\le c \left(\int_\Omega |\nabla \bar{y}|^{\frac{nq}{n-q}} dx \right)^{\frac{n-q}{nq}} \|z\|_{W_0^{1,\frac{nq}{n-q}}(\Omega)} \\ &\le c \|\bar{y}\|_{W^{2,q}(\Omega)} \|z\|_{W_0^{1,\frac{nq}{n-q}}(\Omega)}, \end{aligned}$$

where we have used that $z \in W_0^{1,\frac{nq}{n-q}}(\Omega)$, which is a consequence of the embedding $L^q(\Omega) \subset W^{-1,\frac{nq}{n-q}}(\Omega)$ along with Theorem 2.10. Notice that we have assumed $q > n/2$. This inequality is equivalent to $nq/(n-q) > n$ and is also behind the estimate of the integral containing $\nabla \bar{y}$. □

Remark 2.12. If $q = 2$, then Theorem 2.11 remains true for $n = 2$ or $n = 3$ if we replace the $C^{1,1}$-regularity of Γ by the convexity of Ω. This is a consequence of the H^2-regularity for the elliptic problems in convex domains; see Grisvard [20].

3. The control problem.

Associated to the state equation (1.1), we introduce the control problem

$$\text{(P)} \qquad \begin{cases} \min J(u) = \displaystyle\int_\Omega L(x, y_u(x), u(x))\, dx, \\ u \in L^\infty(\Omega), \\ \alpha(x) \le u(x) \le \beta(x) \text{ for a.e. } x \in \Omega, \end{cases}$$

where $L : \Omega \times (\mathbb{R} \times \mathbb{R}) \to \mathbb{R}$ is a Carathéodory function, $p > n$, and $\alpha, \beta \in L^\infty(\Omega)$, with $\beta(x) \ge \alpha(x)$ for a.e. $x \in \Omega$. A standard example for the choice of L is the quadratic function

$$L(x, y, u) = (y - y_d(x))^2 + \frac{N}{2} u^2,$$

where $y_d \in L^q(\Omega)$ is given fixed.

First of all, we study the existence of a solution for problem (P).

Theorem 3.1. *Let us assume that* (A1) *and* (A2) *hold. We also suppose that L is convex with respect to u and, for any $M > 0$, there exists a function $\psi_M \in L^1(\Omega)$ such that*

$$|L(x, y, u)| \leq \psi_M(x) \quad \textit{for a.e. } x \in \Omega \quad \textit{and} \quad |y|, |u| \leq M.$$

Then (P) *has at least one optimal solution $\bar{u}$.*

Proof. Let $\{u_k\}_{k=1}^{\infty} \subset L^{\infty}(\Omega)$ be a minimizing sequence for (P). Since $\{u_k\}_{k=1}^{\infty}$ is bounded in $L^{\infty}(\Omega) \subset W^{-1,p}(\Omega)$, Theorem 2.4 implies that $\{y_{u_k}\}_{k=1}^{\infty}$ is bounded in $W_0^{1,p}(\Omega)$ and, taking a subsequence, denoted in the same way, we get $u_k \rightharpoonup \bar{u}$ weakly* in $L^{\infty}(\Omega)$, and hence strongly in $W^{-1,p}(\Omega)$. Therefore, $y_{u_k} \to \bar{y}_u$ in $W_0^{1,p}(\Omega)$. Moreover, it is obvious that $\alpha \leq \bar{u} \leq \beta$, and hence $\bar{u}$ is a feasible control for (P). Let us denote by $\bar{y}$ the state associated to $\bar{u}$. Now we prove that $\bar{u}$ is a solution of (P). It is enough to use the convexity of L with respect to u along with the continuity with respect to (y, u) and the Lebesgue dominated convergence theorem as follows:

$$\begin{aligned} J(\bar{u}) = \int_\Omega L(x, \bar{y}(x), \bar{u}(x))\, dx &\leq \liminf_{k\to\infty} \int_\Omega L(x, \bar{y}(x), u_k(x))\, dx \\ \leq \limsup_{k\to\infty} \int_\Omega &|L(x, \bar{y}(x), u_k(x)) - L(x, y_{u_k}(x), u_k(x))|\, dx \\ + \limsup_{k\to\infty} \int_\Omega L(x, y_{u_k}(x), u_k(x))\, dx &= \lim_{k\to\infty} J(u_k) = \inf\,(\mathrm{P}). \qquad \square \end{aligned}$$

Our next goal is to derive the first-order optimality conditions. We get the optimality conditions satisfied by $\bar{u}$ from the standard variational inequality $J'(\bar{u})(u - \bar{u}) \geq 0$ for any feasible control u. To argue in this way, we need the differentiability of J, which requires the differentiability of L with respect to u and y. Since we also wish to derive second-order optimality conditions, we require the existence of the second-order derivatives of L. More precisely, our assumption is the following.

(A6) $L : \Omega \times (\mathbb{R} \times \mathbb{R}) \longrightarrow \mathbb{R}$ is a Carathéodory function of class C^2 with respect to the last two variables and, for all $M > 0$, there exist a constant $C_{L,M} > 0$ and functions $\psi_{u,M} \in L^2(\Omega)$ and $\psi_{y,M} \in L^q(\Omega)$, such that

$$\left|\frac{\partial L}{\partial u}(x, y, u)\right| \leq \psi_{u,M}(x), \quad \left|\frac{\partial L}{\partial y}(x, y, u)\right| \leq \psi_{y,M}(x), \quad \|D^2_{(y,u)} L(x, y, u)\| \leq C_{L,M},$$

$$\|D^2_{(y,u)} L(x, y_2, u_2) - D^2_{(y,u)} L(x, y_1, u_1)\| \leq C_{L,M}(|y_2 - y_1| + |u_2 - u_1|)$$

for a.e. $x \in \Omega$ and $|y|, |y_i|, |u|, |u_i| \leq M$, $i = 1, 2$, where $D^2_{(y,u)} L$ denotes the second derivative of L with respect to (y, u), i.e., the associated Hessian matrix.

By applying the chain rule and introducing the adjoint state as usual, an elementary calculus leads to the following result.

Theorem 3.2. *Let us assume that $a : \bar{\Omega} \times \mathbb{R} \mapsto \mathbb{R}$ is continuous, Γ is of class C^1, and* (A1), (A2), (A4), *and* (A6) *hold. Then the function $J : L^{\infty}(\Omega) \to \mathbb{R}$ is of class C^2. Moreover, for every $u, v, v_1, v_2 \in L^{\infty}(\Omega)$, we have*

$$J'(u)v = \int_\Omega \left(\frac{\partial L}{\partial u}(x, y_u, u) + \varphi_u\right) v\, dx \tag{3.1}$$

and

$$\begin{aligned} J''(u)v_1v_2 = \int_\Omega \Big\{ & \frac{\partial^2 L}{\partial y^2}(x,y_u,u)z_{v_1}z_{v_2} + \frac{\partial^2 L}{\partial y \partial u}(x,y_u,u)(z_{v_1}v_2 + z_{v_2}v_1) \\ & + \frac{\partial^2 L}{\partial u^2}(x,y_u,u)v_1v_2 - \varphi_u \frac{\partial^2 f}{\partial y^2}(x,y_u)z_{v_1}z_{v_2} \\ & - \nabla\varphi_u \left[\frac{\partial a}{\partial y}(x,y_u)(z_{v_1}\nabla z_{v_2} + \nabla z_{v_1} z_{v_2}) + \frac{\partial^2 a}{\partial y^2}(x,y)z_{v_1}z_{v_2}\nabla y_u \right] \Big\} \, dx, \end{aligned} \tag{3.2}$$

where $\varphi_u \in W_0^{1,p}(\Omega)$ *is the unique solution of the problem*

$$\left\{ \begin{aligned} -\operatorname{div}[a(x,y_u)\nabla\varphi] + \frac{\partial a}{\partial y}(x,y_u)\nabla y_u \cdot \nabla\varphi + \frac{\partial f}{\partial y}(x,y_u)\varphi = \frac{\partial L}{\partial y}(x,y_u,u) \ \textit{in } \Omega, \\ \varphi = 0 \ \textit{on } \Gamma, \end{aligned} \right. \tag{3.3}$$

where $z_{v_i} = G'(u)v_i$ *is the solution of* (2.21) *for* $y = y_u$ *and* $v = v_i$, $i = 1, 2$.

Proof. The only delicate point in the proof of the previous theorem is the existence and uniqueness of a solution of the adjoint state equation (3.3). To prove this, let us consider the linear operator $T \in \mathcal{L}(H_0^1(\Omega), H^{-1}(\Omega))$ given by

$$Tz = -\operatorname{div}\left[a(x,y)\nabla z + \frac{\partial a}{\partial y}(x,y)z\,\nabla y \right] + \frac{\partial f}{\partial y}(x,y)\,z.$$

According to Remark 2.8, T is an isomorphism and its adjoint operator is also an isomorphism $T^* \in \mathcal{L}(H_0^1(\Omega), H^{-1}(\Omega))$ given by

$$T^*\varphi = -\operatorname{div}[a(x,y_u)\nabla\varphi] + \frac{\partial a}{\partial y}(x,y_u)\nabla y_u \cdot \nabla\varphi + \frac{\partial f}{\partial y}(x,y_u)\varphi.$$

This is exactly equivalent to the well-posedness of the adjoint equation (3.3) in $H_0^1(\Omega)$. Finally, Theorems 2.2 and 2.4 along with assumption (A6) imply that the adjoint state φ belongs to the space $W_0^{1,p}(\Omega)$, as claimed in the theorem, provided that the term

$$\frac{\partial a}{\partial y}(x,y_u)\nabla y_u \cdot \nabla\varphi$$

belongs to $W^{-1,p}(\Omega)$. Let us prove this fact. Thanks to the boundedness of y_u and the assumption (A4), it is enough to prove that $\nabla y_u \cdot \nabla\varphi \in L^r(\Omega) \subset W^{-1,p}(\Omega)$ holds for some r large enough. By using that $\nabla y_u \in L^p(\Omega)$, $\nabla\varphi \in L^2(\Omega)$ and invoking the Hölder inequality, we get that $\nabla y_u \cdot \nabla\varphi \in L^{2p/(p+2)}(\Omega)$. For $n = 2$, $L^{2p/(p+2)}(\Omega) \subset W^{-1,p}(\Omega)$. Let us consider the case $n > 2$. In this case, we have

$$L^{2p/(p+2)}(\Omega) \subset W^{-1,r}(\Omega), \quad \text{with } r = \frac{2pn}{p(n-2)+2n}.$$

Therefore it turns out that $\varphi \in W_0^{1,\sigma}(\Omega)$, with $\sigma = \min\{p, r\}$. If $\sigma = p$, then the proof is complete. If it is not true, then let us notice that

$$r = 2 + \varepsilon, \quad \text{with } \varepsilon = \frac{4(p-n)}{p(n-2)+2n}.$$

The proof proceeds by induction: For $k \geq 1$, we assume that $\varphi \in W_0^{1,2+k\varepsilon}(\Omega)$ and then we prove that $\varphi \in W_0^{1,\sigma}(\Omega)$, with $\sigma = \min\{p, 2+(k+1)\varepsilon\}$. Consequently, for k large enough, we have that $\sigma = p$. By using the embedding of Sobolev spaces in L^r spaces and after performing some obvious computations, we get that

$$\nabla y_u \in L^p(\Omega) \text{ and } \nabla\varphi \in L^{2+k\varepsilon}(\Omega) \Rightarrow \nabla y_u \cdot \nabla\varphi \in W^{-1,r}(\Omega),$$

with

$$r = \frac{pn(2+k\varepsilon)}{p[n-(2+k\varepsilon)] + (2+k\varepsilon)n}.$$

We have to prove that $r - (2+k\varepsilon) \geq \varepsilon$, which is equivalent to

$$\frac{(p-n)(2+k\varepsilon)^2}{p[n-(2+k\varepsilon)] + (2+k\varepsilon)n} \geq \varepsilon.$$

From the definition of ε, we obtain that the previous inequality is equivalent to

$$(p-n)(2+k\varepsilon)^2 \geq \frac{4(p-n)}{p(n-2)+2n}\{p[n-(2+k\varepsilon)] + (2+k\varepsilon)n\}$$

if and only if

$$[p(n-2)+2n](2+k\varepsilon)^2 \geq 4\{p[n-(2+k\varepsilon)] + (2+k\varepsilon)n\}.$$

Let us set for every $p \geq n$

$$\rho(p) = [p(n-2)+2n](2+k\varepsilon(p))^2,\ \mu(p) = 4\{p[n-(2+k\varepsilon(p))] + (2+k\varepsilon(p))n\}$$

and

$$\varepsilon(p) = \frac{4(p-n)}{p(n-2)+2n}.$$

Using that $\varepsilon(n) = 0$, we get that $\rho(n) = 4n^2 = \mu(n)$. If we prove that $\rho'(p) > \mu'(p)$ for every $p > n$, then the inequality $\rho(p) > \mu(p)$ will be true for all $p > n$ and the proof of the theorem is concluded. Using that $\varepsilon'(p) > 0$ and $\varepsilon(p) > 0$ for $p > n$, we get

$$\rho'(p) = (n-2)(2+k\varepsilon(p))^2 + 2k[p(n-2)+2n](2+k\varepsilon(p))\varepsilon'(p) > 4(n-2)$$

and

$$\begin{aligned}\mu'(p) &= 4(n-2-k\varepsilon(p)) + 4(-kp\varepsilon'(p) + kn\varepsilon'(p)) \\ &= 4(n-2) - 4k[\varepsilon(p) + (p-n)\varepsilon'(p)] < 4(n-2),\end{aligned}$$

which leads to the desired result. □

Remark 3.3. By using the expression given by (3.2) for $J''(u)$, it is obvious that $J''(u)$ can be extended to a continuous bilinear form $J''(u) : L^2(\Omega) \times L^2(\Omega) \longrightarrow \mathbb{R}$.

By using the inequality $J'(\bar{u})(u-\bar{u}) \geq 0$ and the differentiability of J given by (3.1) and (3.3) we deduce the first-order optimality conditions.

THEOREM 3.4. *Under the assumptions of Theorem* 3.2, *if $\bar u$ is a local minimum of* (P), *then there exists $\bar\varphi \in W_0^{1,p}(\Omega)$ such that*

$$\text{(3.4)}\qquad \left\{\begin{array}{r} -\mathrm{div}\,[a(x,\bar y)\nabla\bar\varphi] + \dfrac{\partial a}{\partial y}(x,\bar y)\nabla\bar y\cdot\nabla\bar\varphi + \dfrac{\partial f}{\partial y}(x,\bar y)\bar\varphi = \dfrac{\partial L}{\partial y}(x,\bar y,\bar u) \ \textit{in}\ \Omega,\\ \bar\varphi = 0\ \textit{on}\ \Gamma,\end{array}\right.$$

$$\text{(3.5)}\qquad \int_\Omega\left(\frac{\partial L}{\partial u}(x,\bar y(x),\bar u(x)) + \bar\varphi(x)\right)(u(x)-\bar u(x))\,dx \ge 0 \quad \forall\ \alpha\le u\le\beta,$$

where $\bar y$ is the state associated to $\bar u$.

From (3.5) we get as usual

$$\text{(3.6)}\quad \bar u(x) = \left\{\begin{array}{ll}\alpha(x) & \text{if } \bar d(x) > 0,\\ \beta(x) & \text{if } \bar d(x) < 0\end{array}\right. \quad\text{and}\quad \bar d(x) = \left\{\begin{array}{ll}\ge 0 & \text{if } \bar u(x) = \alpha(x),\\ \le 0 & \text{if } \bar u(x) = \beta(x),\\ = 0 & \text{if } \alpha(x) < \bar u(x) < \beta(x)\end{array}\right.$$

for almost all $x\in\Omega$, where

$$\text{(3.7)}\qquad \bar d(x) = \frac{\partial L}{\partial u}(x,\bar y(x),\bar u(x)) + \bar\varphi(x).$$

We finish this section by studying the regularity of the optimal solutions of (P).

THEOREM 3.5. *Under the assumptions of Theorem* 3.4 *and assuming that*

$$\text{(3.8)}\qquad \frac{\partial L}{\partial u} : \bar\Omega\times(\mathbb{R}\times\mathbb{R})\to\mathbb{R}\ \textit{is continuous,}$$

$$\text{(3.9)}\quad \exists\Lambda_L > 0\ \textit{such that}\ \frac{\partial^2 L}{\partial u^2}(x,y,u)\ge\Lambda_L\ \ \textit{for a.e. } x\in\Omega\ \textit{and}\ \forall y,u\in\mathbb{R}^2,$$

then the equation

$$\text{(3.10)}\qquad \frac{\partial L}{\partial u}(x,\bar y(x),t) + \bar\varphi(x) = 0$$

has a unique solution $\bar t = \bar s(x)$ for every $x\in\bar\Omega$. The function $\bar s:\bar\Omega\to\mathbb{R}$ is continuous and is related to $\bar u$ by the formula

$$\text{(3.11)}\qquad \bar u(x) = \mathit{Proj}_{[\alpha(x),\beta(x)]}(\bar s(x)) = \max\{\min\{\beta(x),\bar s(x)\},\alpha(x)\}.$$

Moreover, if α,β are contained in $C(\bar\Omega)$, then $\bar u$ belongs to $C(\bar\Omega)$, too. Finally, if Γ is $C^{1,1}$, (A3) *holds, $q>n$ is taken in the assumptions* (A2) *and* (A6), *$\alpha,\beta\in C^{0,1}(\bar\Omega)$, and for every $M>0$ there exists a constant $C_{L,M}>0$ such that*

$$\text{(3.12)}\quad \left|\frac{\partial L}{\partial u}(x_2,y,u) - \frac{\partial L}{\partial u}(x_1,y,u)\right| \le C_{L,M}|x_2-x_1|\quad \forall x_i\in\Omega\ \textit{and}\ \forall|y|,|u|\le M,$$

then $\bar s,\bar u\in C^{0,1}(\bar\Omega)$.

Proof. Given $x\in\bar\Omega$, let us define the function $g:\mathbb{R}\to\mathbb{R}$ by

$$g(t) = \frac{\partial L}{\partial u}(x,\bar y(x),t) + \bar\varphi(x).$$

Then g is of class C^1 and from (3.9) we know that it is strictly increasing and

$$\lim_{t\to-\infty} g(t) = -\infty \quad \text{and} \quad \lim_{t\to+\infty} g(t) = +\infty.$$

Therefore, there exists a unique element $\bar{t} \in \mathbb{R}$ such that $g(\bar{t}) = 0$.

Taking $\bar{d}$ as defined by (3.7) and using (3.6) along with the strict monotonicity of $(\partial L/\partial u)$ with respect to the third variable, we obtain

$$\begin{cases} \text{if } \bar{d}(x) > 0, \text{ then } \alpha(x) = \bar{u}(x) > \bar{s}(x), \\ \text{if } \bar{d}(x) < 0, \text{ then } \beta(x) = \bar{u}(x) < \bar{s}(x), \\ \text{if } \bar{d}(x) = 0, \text{ then } \bar{u}(x) = \bar{s}(x), \end{cases}$$

which implies (3.11).

Let us prove that $\bar{s}$ is a bounded function. By using the mean value theorem along with (3.8), (3.9), and (3.10), we get

$$\Lambda_L|\bar{s}(x)| \leq \left|\frac{\partial L}{\partial u}(x,\bar{y}(x),\bar{s}(x)) - \frac{\partial L}{\partial u}(x,\bar{y}(x),0)\right| = \left|\bar{\varphi}(x) + \frac{\partial L}{\partial u}(x,\bar{y}(x),0)\right|,$$

and hence

$$|\bar{s}(x)| \leq \frac{1}{\Lambda_L} \max_{x\in\bar{\Omega}} \left|\bar{\varphi}(x) + \frac{\partial L}{\partial u}(x,\bar{y}(x),0)\right| < \infty.$$

The continuity of $\bar{s}$ at every point $x \in \bar{\Omega}$ follows easily from the continuity of $\bar{y}$ and $(\partial L/\partial u)$ by using the inequality

$$\Lambda_L|\bar{s}(x) - \bar{s}(x')| \leq \left|\frac{\partial L}{\partial u}(x',\bar{y}(x'),\bar{s}(x)) - \frac{\partial L}{\partial u}(x',\bar{y}(x'),\bar{s}(x'))\right|$$

$$\leq |\bar{\varphi}(x') - \bar{\varphi}(x)| + \left|\frac{\partial L}{\partial u}(x',\bar{y}(x'),\bar{s}(x)) - \frac{\partial L}{\partial u}(x,\bar{y}(x),\bar{s}(x))\right|. \tag{3.13}$$

If $\alpha, \beta \in C(\bar{\Omega})$, then the identity (3.11) and the continuity of $\bar{s}$ imply the continuity of $\bar{u}$ in $\bar{\Omega}$.

Finally, if Γ is $C^{1,1}$ and (A3) and (A6) hold with $q > n$, then $\bar{y}, \bar{\varphi} \in W^{2,q}(\Omega) \subset C^{0,1}(\Omega)$. Then we can get from (3.13), the boundedness of $\bar{s}$, and (3.12) that $\bar{s} \in C^{0,1}(\bar{\Omega})$. Once again, (3.11) allows us to conclude that $\bar{u} \in C^{0,1}(\bar{\Omega})$, assuming that α and β are also Lipschitz in $\bar{\Omega}$. Indeed, it is enough to realize that

$$\begin{aligned} |\bar{u}(x_2) - \bar{u}(x_1)| &\leq \max\{|\beta(x_2) - \beta(x_1)|, |\alpha(x_2) - \alpha(x_1)|, |\bar{s}(x_2) - \bar{s}(x_1)|\} \\ &\leq \max\{L_\beta, L_\alpha, L_{\bar{s}}\}|x_2 - x_1|, \end{aligned}$$

where L_β, L_α, and $L_{\bar{s}}$ are the Lipschitz constants of α, β, and $\bar{s}$, respectively. □

4. Pontryagin's principle. The goal of this section is to derive the Pontryagin principle satisfied by a local solution of (P). We need this principle for our second-order analysis. There is already an extensive list of contributions about Pontryagin's principle, but none of them was devoted to quasilinear equations of nonmonotone type. This lack of monotonicity requires an adaptation of the usual proofs to overcome this difficulty. For this purpose, we will make the following assumption.

(A7) $L : \Omega \times (\mathbb{R} \times \mathbb{R}) \longrightarrow \mathbb{R}$ is a Carathéodory function of class C^1 with respect to the second variable and, for all $M > 0$, there exists a function $\psi_M \in L^q(\Omega)$, with $q \geq pn/(p+n)$, such that

$$\left|\frac{\partial L}{\partial y}(x,y,u)\right| \leq \psi_M(x) \text{ for a.e. } x \in \Omega,\ |u| \leq M, \text{ and } |y| \leq M.$$

Associated with the control problem (P), we define the Hamiltonian as usual by

$$H(x,y,u,\varphi) = L(x,y,u) + \varphi[u - f(x,y)].$$

The Pontryagin principle is formulated as follows.

THEOREM 4.1. *Let $\bar{u}$ be a local solution of* (P)*. We assume that $a : \bar{\Omega} \times \mathbb{R} \mapsto \mathbb{R}$ is continuous, Γ is of class C^1, and* (A1)*,* (A2)*,* (A4)*, and* (A7) *hold. Then there exists $\bar{\varphi} \in W_0^{1,p}(\Omega)$ satisfying* (3.4) *and*

$$H(x,\bar{y}(x),\bar{u}(x),\bar{\varphi}(x)) = \min_{s \in [\alpha_{\varepsilon_{\bar{u}}}(x),\beta_{\varepsilon_{\bar{u}}}(x)]} H(x,\bar{y}(x),s,\bar{\varphi}(x)) \quad \textit{for a.e. } x \in \Omega, \tag{4.1}$$

where

$$\alpha_{\varepsilon_{\bar{u}}}(x) = \max\{\alpha(x), \bar{u}(x) - \varepsilon_{\bar{u}}\} \quad \textit{and} \quad \beta_{\varepsilon_{\bar{u}}}(x) = \min\{\beta(x), \bar{u}(x) + \varepsilon_{\bar{u}}\},$$

$\varepsilon_{\bar{u}} > 0$ is the radius of the $L^\infty(\Omega)$ ball where J achieves the minimum value at $\bar{u}$ among all feasible controls.

Relation (4.1) is an immediate consequence of (3.5) if L is convex with respect to the third variable, but this assumption is not made in the above theorem. To prove (4.1), we will use the following lemma whose proof can be found in [13, Lemma 4.3].

LEMMA 4.2. *For every $0 < \rho < 1$, there exists a sequence of Lebesgue measurable sets $\{E_k\}_{k=1}^\infty \subset \Omega$ such that*

$$|E_k| = \rho|\Omega| \quad \textit{and} \quad \frac{1}{\rho}\chi_{E_k} \rightharpoonup 1 \quad \textit{in } L^\infty(\Omega) \quad \textit{weakly}^\star, \tag{4.2}$$

where $|\cdot|$ denotes the Lebesgue measure.

PROPOSITION 4.3. *Under the assumptions of Theorem* 4.1*, for any $u \in L^\infty(\Omega)$ there exist a number $0 < \hat{\rho} < 1$ and measurable sets $E_\rho \subset \Omega$, with $|E_\rho| = \rho|\Omega|$ for all $0 < \rho < \hat{\rho}$, that have the following properties: If we define*

$$u_\rho(x) = \begin{cases} \bar{u}(x) & \textit{if } x \in \Omega \setminus E_\rho, \\ u(x) & \textit{if } x \in E_\rho, \end{cases}$$

then

$$y_\rho = \bar{y} + \rho z + r_\rho, \quad \lim_{\rho \searrow 0} \frac{1}{\rho}\|r_\rho\|_{W_0^{1,p}(\Omega)} = 0, \tag{4.3}$$

$$J(u_\rho) = J(\bar{u}) + \rho z^0 + r_\rho^0, \quad \lim_{\rho \searrow 0} \frac{1}{\rho}|r_\rho^0| = 0 \tag{4.4}$$

hold true, where $\bar{y}$ and y_ρ are the states associated to $\bar{u}$ and u_ρ, respectively, z is the unique element of $W_0^{1,p}(\Omega)$ satisfying the linearized equation

$$\operatorname{div}\left[a(x,\bar{y})\nabla z + \frac{\partial a}{\partial y}(x,\bar{y})z\,\nabla\bar{y}\right] + \frac{\partial f}{\partial y}(x,\bar{y})\,z = u - \bar{u} \quad \textit{in } \Omega, \tag{4.5}$$

and

$$z^0 = \int_\Omega \left\{ \frac{\partial L}{\partial y}(x, \bar{y}(x), \bar{u}(x))z(x) + L(x, \bar{y}(x), u(x)) - L(x, \bar{y}(x), \bar{u}(x)) \right\} dx. \tag{4.6}$$

Proof. Let us define the function $g \in L^1(\Omega)$ by

$$g(x) = L(x, \bar{y}(x), u(x)) - L(x, \bar{y}(x), \bar{u}(x)).$$

Given $\rho \in (0, 1)$, we take a sequence $\{E_k\}_{k=1}^\infty$ as in Lemma 4.2. Since $L^\infty(\Omega)$ is compactly embedded in $W^{-1,p}(\Omega)$, there exists k_ρ such that

$$\left| \int_\Omega \left(1 - \frac{1}{\rho}\chi_{E_k}(x)\right) g(x)\, dx \right| + \left\| \left(1 - \frac{1}{\rho}\chi_{E_k}\right)(u - \bar{u}) \right\|_{W^{-1,p}(\Omega)} < \rho \ \ \forall k \geq k_\rho. \tag{4.7}$$

Let us denote $E_\rho = E_{k_\rho}$. Let us introduce $z_\rho = (y_\rho - \bar{y})/\rho$. By subtracting the equations satisfied by y_ρ and $\bar{y}$ and dividing by ρ we get

$$-\text{div} \left[a(x, \bar{y})\nabla z_\rho + \frac{a(x, y_\rho) - a(x, \bar{y})}{\rho} \nabla y_\rho \right] + \frac{f(x, y_\rho) - f(x, \bar{y})}{\rho} = \frac{u_\rho - \bar{u}}{\rho} \ \text{ in } \Omega.$$

Now setting

$$a_\rho(x) = \int_0^1 \frac{\partial a}{\partial y}(x, \bar{y}(x) + \theta(y_\rho(x) - \bar{y}(x)))\, d\theta,$$
$$f_\rho(x) = \int_0^1 \frac{\partial f}{\partial y}(x, \bar{y}(x) + \theta(y_\rho(x) - \bar{y}(x)))\, d\theta$$

we deduce from the above identity

$$-\text{div}\, [a(x, \bar{y})\nabla z_\rho + a_\rho(x) z_\rho \nabla y_\rho] + f_\rho(x) z_\rho = \frac{1}{\rho}\chi_{E_\rho}(u - \bar{u}) \ \text{ in } \Omega. \tag{4.8}$$

Let us define $T, T_\rho : W_0^{1,p}(\Omega) \mapsto W^{-1,p}(\Omega)$ by

$$T\xi = -\text{div} \left[a(x, \bar{y})\nabla \xi + \frac{\partial a}{\partial y}(x, \bar{y})\xi \nabla \bar{y} \right] + \frac{\partial f}{\partial y}(x, \bar{y})\xi,$$
$$T_\rho \xi = -\text{div}\, [a(x, \bar{y})\nabla \xi + a_\rho(x)\xi \nabla y_\rho] + f_\rho(x)\xi.$$

Since $y_\rho \to \bar{y}$ in $W_0^{1,p}(\Omega) \subset C(\bar{\Omega})$, we deduce from our assumptions on a and f that

$$a_\rho(x) \to \frac{\partial a}{\partial y}(x, \bar{y}(x)) \ \text{ and } \ f_\rho(x) \to \frac{\partial f}{\partial y}(x, \bar{y}(x)) \text{ uniformly in } \bar{\Omega}, \tag{4.9}$$

and consequently

$$\|T_\rho - T\|_{\mathcal{L}(W_0^{1,p}(\Omega), W^{-1,p}(\Omega))} \leq C \left\{ \|y_\rho - \bar{y}\|_{W_0^{1,p}(\Omega)} \right.$$
$$\left. + \|a_\rho(x) - \frac{\partial a}{\partial y}(x, \bar{y}(x))\|_{C(\bar{\Omega})} + \|f_\rho(x) - \frac{\partial f}{\partial y}(x, \bar{y}(x))\|_{C(\bar{\Omega})} \right\} \to 0. \tag{4.10}$$

Since T is an isomorphism, by taking $\hat{\rho}$ small enough, we have that T_ρ is also an isomorphism and $T_\rho^{-1} \to T^{-1}$ in $\mathcal{L}(W^{-1,p}(\Omega), W_0^{1,p}(\Omega))$ too. Taking into account (4.7), we obtain

$$\|z - z_\rho\|_{W_0^{1,p}(\Omega)} = \left\| T^{-1}(u-\bar{u}) - T_\rho^{-1}\left[\frac{1}{\rho}\chi_{E_\rho}(u-\bar{u})\right]\right\|_{W_0^{1,p}(\Omega)}$$

$$\leq \left\| T_\rho^{-1}\left[\left(1 - \frac{1}{\rho}\chi_{E_\rho}\right)(u-\bar{u})\right]\right\|_{W_0^{1,p}(\Omega)} + \|(T^{-1} - T_\rho^{-1})(u-\bar{u})\|_{W_0^{1,p}(\Omega)}$$

$$\leq C\left\|\left(1 - \frac{1}{\rho}\chi_{E_\rho}\right)(u-\bar{u})\right\|_{W^{-1,p}(\Omega)}$$

$$+ \|T^{-1} - T_\rho^{-1}\|_{\mathcal{L}(W_0^{1,p}(\Omega),W^{-1,p}(\Omega))}\|u-\bar{u}\|_{W^{-1,p}(\Omega)} \to 0.$$

Now it is enough to notice that, by definition of z_ρ and the convergence $z_\rho \to z$ in $W_0^{1,p}(\Omega)$, we have

$$\varepsilon_\rho = \frac{y_\rho - \bar{y}}{\rho} - z \to 0,$$

and hence $y_\rho = \bar{y} + \rho z + \rho\varepsilon_\rho$. By putting $r_\rho = \rho\varepsilon_\rho$ we get (4.3). Finally, let us prove (4.4). Similarly to the definitions of a_ρ and f_ρ, we introduce

$$L_\rho(x) = \int_0^1 \frac{\partial L}{\partial y}(x, \bar{y}(x) + \theta(y_\rho(x) - \bar{y}(x)), u_\rho(x))\, d\theta.$$

Then we have

$$\frac{J(u_\rho) - J(\bar{u})}{\rho} = \int_\Omega \frac{L(x, y_\rho(x), u_\rho(x)) - L(x, \bar{y}(x), \bar{u}(x))}{\rho}\, dx$$

$$= \int_\Omega \frac{L(x, y_\rho(x), u_\rho(x)) - L(x, \bar{y}(x), u_\rho(x))}{\rho}\, dx$$

$$+ \int_\Omega \frac{L(x, \bar{y}(x), u_\rho(x)) - L(x, \bar{y}(x), \bar{u}(x))}{\rho}\, dx$$

$$= \int_\Omega L_\rho(x) z_\rho(x)\, dx + \int_\Omega \frac{1}{\rho}\chi_{E_\rho}(x)[L(x, \bar{y}(x), u(x)) - L(x, \bar{y}(x), \bar{u}(x))]dx$$

$$\to \int_\Omega \frac{\partial L}{\partial y}(x, \bar{y}(x), \bar{u}(x)) z(x)\, dx + \int_\Omega [L(x, \bar{y}(x), u(x)) - L(x, \bar{y}(x), \bar{u}(x))]dx = z^0,$$

which implies (4.4). $\square$

Proof of Theorem 4.1. Since $\bar{u}$ is a local solution of (P), there exists $\varepsilon_{\bar{u}} > 0$ such that J achieves the minimum at $\bar{u}$ among all feasible controls of $\bar{B}_{L^\infty(\Omega)}(\bar{u}, \varepsilon_{\bar{u}})$. Let us take $u \in B_{L^\infty(\Omega)}(\bar{u}, \varepsilon_{\bar{u}})$ with $\alpha(x) \leq u(x) \leq \beta(x)$ a.e. $x \in \Omega$. Following Proposition 4.3, we consider the sets $\{E_\rho\}_{\rho > 0}\}$ such that (4.3) and (4.4) hold. Then $u_\rho \in B_{L^\infty(\Omega)}(\bar{u}, \varepsilon_{\bar{u}})$ and therefore (4.4) leads to

$$0 \leq \lim_{\rho \searrow 0} \frac{J(u_\rho) - J(\bar{u})}{\rho} = z^0.$$

By using (4.5) and the adjoint state given by (3.4), we get from the previous inequality after an integration by parts

$$0 \le \int_\Omega \{\bar\varphi(x)(u(x)-\bar u(x)) + L(x,\bar y(x),u(x)) - L(x,\bar y(x),\bar u(x))\}\,dx$$

$$= \int_\Omega [H(x,\bar y(x),u(x),\bar\varphi(x)) - H(x,\bar y(x),\bar u(x),\bar\varphi(x))]dx. \tag{4.11}$$

Since u is an arbitrary feasible control in the ball $B_{L^\infty(\Omega)}(\bar u,\varepsilon_{\bar u})$, taking into account the definitions of $\alpha_{\varepsilon_{\bar u}}$ and $\beta_{\varepsilon_{\bar u}}$ given in the statement of Theorem 4.1, we deduce from (4.11)

$$\int_\Omega H(x,\bar y(x),\bar u(x),\bar\varphi(x))\,dx = \min_{\alpha_{\varepsilon_{\bar u}}\le u\le\beta_{\varepsilon_{\bar u}}} \int_\Omega [H(x,\bar y(x),u(x),\bar\varphi(x))\,dx. \tag{4.12}$$

To conclude the proof, we will show that (4.12) implies (4.1). Let the sequence $\{q_j\}_{j=1}^\infty$ exhaust the rational numbers contained in $[0,1]$. For every j we set $u_j = q_j\alpha_{\varepsilon_{\bar u}} + (1-q_j)\beta_{\varepsilon_{\bar u}}$. Then every function u_j belongs to $L^\infty(\Omega)$ and $\alpha_{\varepsilon_{\bar u}}(x) \le u_j(x) \le \beta_{\varepsilon_{\bar u}}(x)$ for every $x\in\Omega$. Now we introduce functions $F_0, F_j : \Omega \mapsto \mathbb{R}$ by

$$F_0(x) = H(x,\bar y(x),\bar u(x),\bar\varphi(x)) \text{ and } F_j(x) = H(x,\bar y(x),u_j(x),\bar\varphi(x)),\quad j=1,\dots,\infty.$$

Associated to these integrable functions we introduce the set of Lebesgue regular points E_0 and $\{E_j\}_{j=1}^\infty$, which are known to satisfy $|E_j| = |\Omega|$ for $j=0,1,\dots,\infty$, and

$$\lim_{r\searrow 0}\frac{1}{|B_r(x_0)|}\int_{B_r(x_0)} F_j(x)\,dx = F_j(x_0)\ \ \forall x_0\in E_j,\ j=0,1,\dots,\infty, \tag{4.13}$$

where $B_r(x_0)$ is the Euclidean ball in $\mathbb{R}^n$ of center x_0 and radius r. Let us set $E = \cap_{j=0}^\infty E_j$. Then it is obvious that $|E| = |\Omega|$ and (4.13) holds for every $x_0 \in E$. Given $x_0\in E$ and $r>0$ we define

$$u_{j,r}(x) = \begin{cases} \bar u(x) & \text{if } x \notin B_r(x_0),\\ u_j(x) & \text{if } x\in B_r(x_0),\ j=1,\dots,\infty.\end{cases}$$

From (4.12) and the above definition we deduce

$$\int_\Omega H(x,\bar y(x),\bar u(x),\bar\varphi(x))\,dx \le \int_\Omega H(x,\bar y(x),u_{j,r}(x),\bar\varphi(x))\,dx,$$

and therefore

$$\frac{1}{|B_r(x_0)|}\int_{B_r(x_0)} H(x,\bar y(x),\bar u(x),\bar\varphi(x))\,dx,$$
$$\le \frac{1}{|B_r(x_0)|}\int_{B_r(x_0)} H(x,\bar y(x),u_j(x),\bar\varphi(x))\,dx,$$

and passing to the limit when $r\to 0$ we get

$$H(x_0,\bar y(x_0),\bar u(x_0),\bar\varphi(x_0)) \le H(x_0,\bar y(x_0),u_j(x_0),\bar\varphi(x_0)).$$

Since the function $s \to H(x_0,\bar y(x_0),s,\bar\varphi(x_0))$ is continuous and $\{u_j(x_0)\}_{j=1}^\infty$ is dense in $[\alpha_{\varepsilon_{\bar u}}(x_0),\beta_{\varepsilon_{\bar u}}(x_0)]$, we get

$$H(x_0,\bar y(x_0),\bar u(x_0),\bar\varphi(x_0)) \le H(x_0,\bar y(x_0),s,\bar\varphi(x_0))\ \ \forall s\in[\alpha_{\varepsilon_{\bar u}}(x_0),\beta_{\varepsilon_{\bar u}}(x_0)].$$

Finally, (4.1) follows from the previous inequality and the fact that x_0 is an arbitrary point of E. □

Remark 4.4. If we consider that $\bar{u}$ is a global solution or even a local solution of (P) in the sense of the $L^p(\Omega)$ topology, then (4.1) holds with $\varepsilon_{\bar{u}} = 0$. More precisely

$$H(x, \bar{y}(x), \bar{u}(x), \bar{\varphi}(x)) = \min_{s \in [\alpha(x), \beta(x)]} H(x, \bar{y}(x), s, \bar{\varphi}(x)) \text{ for a.e. } x \in \Omega.$$

The proof is the same. The only point we have to address is that the functions u_ρ defined in Proposition 4.3 corresponding to feasible controls u satisfy

$$\begin{aligned}\|u_p - \bar{u}\|_{L^p(\Omega)} = \left(\int_{E_\rho} |u(x) - \bar{u}(x)|^p \, dx \right)^{1/p} &\leq \|u - \bar{u}\|_{L^\infty(\Omega)} |E_\rho|^{1/p} \\ &\leq \|\beta - \alpha\|_{L^\infty(\Omega)} |\Omega|^{1/p} \rho^{1/p}.\end{aligned}$$

Therefore for ρ small enough the functions u_ρ are in the corresponding ball of $L^p(\Omega)$ where $\bar{u}$ is the minimum.

5. Second-order optimality conditions. The goal of this section is to prove first necessary and next sufficient second-order optimality conditions. For it we will assume that (A1), (A2), (A4), and (A6) hold, the function $a : \bar{\Omega} \times \mathbb{R} \longrightarrow \mathbb{R}$ is continuous, and Γ is of class C^1.

If $\bar{u}$ is a feasible control for problem (P) and there exists $\bar{\varphi} \in W_0^{1,p}(\Omega)$ satisfying (3.4) and (3.5), then we introduce the cone of critical directions

$$(5.1) \qquad C_{\bar{u}} = \left\{ h \in L^2(\Omega) : h(x) = \left\{ \begin{array}{ll} \geq 0 & \text{if } \bar{u}(x) = \alpha(x) \\ \leq 0 & \text{if } \bar{u}(x) = \beta(x) \\ = 0 & \text{if } \bar{d}(x) \neq 0 \end{array} \right. \text{ for a.e. } x \in \Omega \right\},$$

where $\bar{d}$ is defined by (3.7). In the previous section, we introduced the Hamiltonian H associated to the control problem. It is easy to check that

$$\frac{\partial H}{\partial u}(x, \bar{y}(x), \bar{u}(x), \bar{\varphi}(x)) = \bar{d}(x).$$

In what follows, we will use the notation

$$\bar{H}_u(x) = \frac{\partial H}{\partial u}(x, \bar{y}(x), \bar{u}(x), \bar{\varphi}(x)) \quad \text{and} \quad \bar{H}_{uu}(x) = \frac{\partial^2 H}{\partial u^2}(x, \bar{y}(x), \bar{u}(x), \bar{\varphi}(x)).$$

Now we prove the necessary second-order optimality conditions.

THEOREM 5.1. *Let us assume that $\bar{u}$ is a local solution of* (P). *Then the following inequalities hold:*

$$(5.2) \qquad \left\{ \begin{array}{l} J''(\bar{u})h^2 \geq 0 \quad \forall h \in C_{\bar{u}}, \\ \bar{H}_{uu}(x) \geq 0 \quad \textit{for a.a. } x \textit{ with } \bar{H}_u(x) = 0. \end{array} \right.$$

Proof. Let us take $h \in C_{\bar{u}}$ arbitrarily and $0 < \varepsilon < \varepsilon_{\bar{u}}$. Then we define

$$h_\varepsilon(x) = \left\{ \begin{array}{l} 0 \text{ if } \quad \alpha(x) < \bar{u}(x) < \alpha(x) + \varepsilon \text{ or } \beta(x) - \varepsilon < \bar{u}(x) < \beta(x), \\ \max\{-\frac{1}{\varepsilon}, \min\{+\frac{1}{\varepsilon}, h(x)\}\} \text{ otherwise.} \end{array} \right.$$

It is clear that $h_\varepsilon \in C_{\bar{u}} \cap L^\infty(\Omega)$ and $h_\varepsilon \to h$ in $L^2(\Omega)$. Moreover, we have

$$\alpha(x) \le \bar{u}(x) + t h_\varepsilon(x) \le \beta(x) \quad \text{for a.e. } x \in \Omega \quad \text{and } 0 \le t < \varepsilon^2.$$

Therefore, if we define $g_\varepsilon : [0, \varepsilon^2] \longrightarrow \mathbb{R}$ by $g_\varepsilon(t) = J(\bar{u} + t h_\varepsilon)$, we have

$$g_\varepsilon(0) = \min_{t \in [0,\varepsilon^2]} g_\varepsilon(t).$$

From our assumptions it is clear that g_ε is a C^2 function. From the fact $h_\varepsilon \in C_{\bar{u}}$ we deduce that

$$g'_\varepsilon(0) = J'(\bar{u}) h_\varepsilon = \int_\Omega \bar{H}_u(x) h_\varepsilon(x)\, dx = 0.$$

Now, an elementary calculus and Theorem 3.2 yield

$$\begin{aligned} 0 \le g''_\varepsilon(0) = J''(\bar{u}) h_\varepsilon^2 = \int_\Omega \Bigg\{ & \frac{\partial^2 L}{\partial y^2}(x,\bar{y},\bar{u}) z_{h_\varepsilon}^2 + 2 \frac{\partial^2 L}{\partial y \partial u}(x,\bar{y},\bar{u}) z_{h_\varepsilon} h_\varepsilon \\ & + \frac{\partial^2 L}{\partial u^2}(x,\bar{y},\bar{u}) h_\varepsilon^2 - \bar{\varphi} \frac{\partial^2 f}{\partial y^2}(x,\bar{y}) z_{h_\varepsilon}^2 \\ & - \nabla \bar{\varphi} \cdot \left[2 \frac{\partial a}{\partial y}(x,\bar{y}) z_{h_\varepsilon} \nabla z_{h_\varepsilon} + \frac{\partial^2 a}{\partial y^2}(x,\bar{y}) z_{h_\varepsilon}^2 \nabla \bar{y} \right] \Bigg\}\, dx, \end{aligned} \tag{5.3}$$

where $z_{h_\varepsilon} \in H_0^1(\Omega)$ is the solution of (2.16) corresponding to $v = h_\varepsilon$. Moreover, the convergence $h_\varepsilon \to h$ in $L^2(\Omega)$ implies that $z_{h_\varepsilon} \to z_h$ in $H_0^1(\Omega)$, where z_h is the solution of (2.16) for $v = h$; see Remark 2.8. Now we estimate the terms of (5.3). Arguing as in Remark 2.8, and taking into account the embedding $H_0^1(\Omega) \subset L^{\frac{2p}{p-2}}(\Omega)$ and assumption (A4), we get

$$\begin{aligned} \int_\Omega \left| \nabla \bar{\varphi}(x) \cdot \frac{\partial a}{\partial y}(x,\bar{y}) z_{h_\varepsilon}(x) \nabla z_{h_\varepsilon}(x) \right| dx &\le D_M \|\nabla \bar{\varphi}\|_{L^p(\Omega)} \|z_{h_\varepsilon}\|_{L^{\frac{2p}{p-2}}(\Omega)} \|\nabla z_{h_\varepsilon}\|_{L^2(\Omega)} \\ &\le C D_M \|\bar{\varphi}\|_{W_0^{1,p}(\Omega)} \|z_{h_\varepsilon}(x)\|_{H_0^1(\Omega)}^2. \end{aligned}$$

Analogously we have

$$\begin{aligned} \int_\Omega \left| \nabla \bar{\varphi}(x) \cdot \frac{\partial^2 a}{\partial y^2}(x,\bar{y}) z_{h_\varepsilon}^2(x) \nabla \bar{y}(x) \right| dx &\le D_M \|\nabla \bar{\varphi}\|_{L^p(\Omega)} \|z_{h_\varepsilon}\|_{L^{\frac{2p}{p-2}}(\Omega)}^2 \|\nabla \bar{y}\|_{L^p(\Omega)} \\ &\le C D_M \|\bar{\varphi}\|_{W_0^{1,p}(\Omega)} \|z_{h_\varepsilon}(x)\|_{H_0^1(\Omega)}^2 \|\bar{y}\|_{W_0^{1,p}(\Omega)}. \end{aligned}$$

The rest of the terms in the integral (5.3) are easy to estimate with the help of assumptions (A4) and (A6). Therefore, we can pass to the limit in (5.3) and deduce

$$0 \le \lim_{\varepsilon \to 0} J''(\bar{u}) h_\varepsilon^2 = J''(\bar{u}) h^2.$$

This proves the first inequality of (5.2). Finally, the second inequality is an obvious consequence of (4.1). Indeed, it is a standard conclusion of (4.1) that

$$\bar{H}_u(x) = \begin{cases} \ge 0 & \text{if } \bar{u}(x) = \alpha(x), \\ \le 0 & \text{if } \bar{u}(x) = \beta(x), \\ = 0 & \text{if } \alpha(x) < \bar{u}(x) < \beta(x) \end{cases} \qquad \text{for a.e. } x \in \Omega$$

and

$$\bar{H}_{uu}(x) \geq 0 \ \text{ if } \ \bar{H}_u(x) = 0 \ \text{ for a.e. } x \in \Omega. \qquad \square$$

Let us consider the Lagrangian function associated to the control problem (P),

$$\mathcal{L} : L^\infty(\Omega) \times W_0^{1,p}(\Omega) \times W_0^{1,p}(\Omega) \longrightarrow \mathbb{R},$$

given by the expression

$$\mathcal{L}(u, y, \varphi) = \mathcal{J}(y, u) + \int_\Omega \{\varphi[u - f(x, y)] - a(x, y)\nabla\varphi \cdot \nabla y\}\, dx$$
$$= \int_\Omega \{H(x, y(x), u(x), \varphi(x)) - a(x, y(x))\nabla\varphi(x) \cdot \nabla y(x)\}\, dx,$$

where we denote

$$\mathcal{J}(y, u) = \int_\Omega L(x, y(x), u(x))\, dx.$$

Defining $\bar{H}_y$, $\bar{H}_{yy}$, and $\bar{H}_{yu}$ similarly to $\bar{H}_u$ and $\bar{H}_{uu}$, after obvious modifications, we can write the first- and second-order derivatives of $\mathcal{L}$ with respect to (y, u) as follows:

$$D_{(y,u)}\mathcal{L}(\bar{u}, \bar{y}, \bar{\varphi})(z, h) = \int_\Omega \left\{\bar{H}_y(x)z(x) + \bar{H}_u(x)h(x)\right\} dx$$
$$- \int_\Omega \nabla\bar{\varphi}(x) \cdot \left\{a(x, \bar{y}(x))\nabla z(x) + \frac{\partial a}{\partial y}(x, \bar{y}(x))z(x)\nabla\bar{y}(x)\right\} dx.$$

If we assume that z is the solution of (2.16) associated to $v = h$, then by using the adjoint state (3.4) we get

$$D_{(y,u)}\mathcal{L}(\bar{u}, \bar{y}, \bar{\varphi})(z, h) = \int_\Omega \bar{H}_u(x)h(x)\, dx. \tag{5.4}$$

Moreover, we find

$$D^2_{(y,u)}\mathcal{L}(\bar{u}, \bar{y}, \bar{\varphi})(z, h)^2 = \int_\Omega \left\{\bar{H}_{yy}(x)z^2(x) + 2\bar{H}_{yu}(x)z(x)h(x) + \bar{H}_{uu}(x)h^2(x)\right\} dx$$
$$- \int_\Omega \nabla\bar{\varphi}(x) \cdot \left\{\frac{\partial^2 a}{\partial y^2}(x, \bar{y}(x))z^2(x)\nabla\bar{y}(x) + 2\frac{\partial a}{\partial y}(x, \bar{y}(x))z(x)\nabla z(x)\right\} dx.$$

Once again if we take z as the solution of (2.16) associated to $v = h$, we deduce from (3.2) that

$$J''(\bar{u})h^2 = D^2_{(y,u)}\mathcal{L}(\bar{u}, \bar{y}, \bar{\varphi})(z, h)^2. \tag{5.5}$$

Therefore the necessary optimality conditions (5.2) can be written as follows:

$$\left\{\begin{array}{l} D^2_{(y,u)}\mathcal{L}(\bar{u}, \bar{y}, \bar{\varphi})(z, h)^2 \geq 0 \ \ \forall (z, h) \in H_0^1(\Omega) \times C_{\bar{u}} \text{ satisfying (2.16)}, \\ \bar{H}_{uu}(x) \geq 0 \ \text{ if } \ \bar{H}_u(x) = 0 \quad \text{for a.e. } x \in \Omega. \end{array}\right. \tag{5.6}$$

We finish this section by establishing the sufficient second-order optimality conditions.

THEOREM 5.2. *Let us assume that $\bar u$ is a feasible control for the problem* (P) *and that there exists $\bar\varphi \in W_0^{1,p}(\Omega)$ satisfying* (3.4) *and* (3.5). *If, in addition, there exist $\mu > 0$ and $\tau > 0$ such that*

$$\begin{aligned} &J''(\bar u)h^2 > 0 \quad \forall h \in C_{\bar u} \setminus \{(0,0)\}, \\ &\bar H_{uu}(x) \geq \mu \quad \textit{if} \quad |\bar H_u(x)| \leq \tau \quad \textit{for a.e. } x \in \Omega, \end{aligned} \tag{5.7}$$

then there exist $\varepsilon > 0$ and $\delta > 0$ such that

$$J(\bar u) + \frac{\delta}{2}\|u - \bar u\|^2_{L^2(\Omega)} \leq J(u) \tag{5.8}$$

for every feasible control $u \in L^\infty(\Omega)$ for (P) *such that $\|u - \bar u\|_{L^\infty(\Omega)} \leq \varepsilon$.*

Remark 5.3. 1. If we compare the first inequality of (5.7) with the analogous inequality of (5.2), we see that the gap is minimal between the necessary and sufficient conditions, as is usual in finite dimensions. However, the second inequality of (5.7) is stronger than the corresponding one of (5.2). This is a consequence of the infinite number of constraints on the control: one constraint for every point of Ω. In general we cannot take $\tau = 0$. The reader is referred to Dunn [17] for a simple example proving the impossibility of taking $\tau = 0$.

2. Let us recall that $\bar H_{uu}(x) = (\partial^2 L/\partial u^2)(x, \bar y(x), \bar u(x))$. Therefore, the second condition of (5.7) is satisfied if we assume that the second derivative of L with respect to u is strictly positive. A standard example is given by the function

$$L(x,y,u) = L_0(x,y) + \frac{N}{2}u^2, \quad \text{with } N > 0.$$

3. The sufficient optimality conditions (5.7) can be written as follows:

$$\begin{aligned} &D^2_{(y,u)}\mathcal{L}(\bar u, \bar y, \bar\varphi)(z,h)^2 > 0 \quad \forall (z,h) \in (H_0^1(\Omega) \times C_{\bar u}) \setminus \{(0,0)\} \text{ verifying (2.16)}, \\ &\bar H_{uu}(x) \geq \mu \quad \text{if} \quad |\bar H_u(x)| \leq \tau \quad \text{for a.e. } x \in \Omega. \end{aligned}$$

Once again this is an obvious consequence of (5.5).

Proof.

Step 1*: Preparations.* We will argue by contradiction. Let us assume that there exists a sequence of feasible controls for (P), $\{u_k\}_{k=1}^\infty \subset L^\infty(\Omega)$, such that

$$\|u_k - \bar u\|_{L^\infty(\Omega)} < \frac{1}{k} \quad \text{and} \quad J(\bar u) + \frac{1}{k}\|u_k - \bar u\|^2_{L^2(\Omega)} > J(u_k). \tag{5.9}$$

Let us define

$$y_k = G(u_k) = y_{u_k},\ \bar y = G(\bar u) = y_{\bar u},\ \rho_k = \|u_k - \bar u\|_{L^2(\Omega)} \ \text{ and } \ v_k = \frac{1}{\rho_k}(u_k - \bar u). \tag{5.10}$$

Then

$$\lim_{k\to\infty} \|y_k - \bar y\|_{W_0^{1,p}(\Omega)} = 0, \ \lim_{k\to\infty} \rho_k = 0 \quad \text{and} \quad \|v_k\|_{L^2(\Omega)} = 1 \ \ \forall k. \tag{5.11}$$

By taking a subsequence, if necessary, we can assume that $v_k \rightharpoonup v$ weakly in $L^2(\Omega)$. We will prove that $v \in C_{\bar u}$. Next, we will use (5.7). In this process we will need the following result:

$$\lim_{k\to\infty} \frac{1}{\rho_k}(y_k - \bar y) = z \quad \text{in} \ \ H_0^1(\Omega), \tag{5.12}$$

where $z \in H_0^1(\Omega)$ is the solution of (2.16) corresponding to the state $\bar{y}$. Let us prove it. We will set $z_k = (y_k - \bar{y})/\rho_k$. By subtracting the state equations satisfied by (y_k, u_k) and $(\bar{y}, \bar{u})$, dividing by ρ_k, and applying the mean value theorem, we get (5.13)

$$-\text{div}\left[a(x,y_k)\nabla z_k + \frac{\partial a}{\partial y}(x,\bar{y}+\theta_k(y_k-\bar{y}))z_k\nabla\bar{y}\right] + \frac{\partial f}{\partial y}(x,\bar{y}+\nu_k(y_k-\bar{y}))z_k = v_k.$$

Taking into account that $z_k \in W_0^{1,p}(\Omega)$, we can multiply (5.13) by z_k and make an integration by parts to get, with the aid of (2.1) and (5.11), that

$$\alpha_0\int_\Omega |\nabla z_k(x)|^2\,dx \leq \int_\Omega a(x,y_k)|\nabla z_k(x)|^2\,dx$$
$$= \int_\Omega \left\{v_k z_k - \frac{\partial f}{\partial y}(x,\bar{y}+\nu_k(y_k-\bar{y}))z_k^2 - \frac{\partial a}{\partial y}(x,\bar{y}+\theta_k(y_k-\bar{y}))z_k\nabla z_k\cdot\nabla\bar{y}\right\}dx$$
$$\leq \|v_k\|_{L^2(\Omega)}\|z_k\|_{L^2(\Omega)} + C\|z_k\|_{L^{\frac{2p}{p-2}}(\Omega)}\|\nabla\bar{y}\|_{L^2(\Omega)}\|\nabla z_k\|_{L^2(\Omega)}.$$

We have used that the term $-\partial f/\partial y\, z_k^2$ is nonpositive. Therefore,

$$\|\nabla z_k\|_{L^2(\Omega)} \leq C\left\{1 + \|z_k\|_{L^{\frac{2p}{p-2}}(\Omega)}\right\}.$$

As in the proof of Theorem 2.7, $\{z_k\}_{k=1}^\infty$ must be bounded in $L^{\frac{2p}{p-2}}(\Omega)$; otherwise we could obtain a nonzero solution of (2.16). Then the above inequality leads to the boundedness of $\{z_k\}_{k=1}^\infty$ in $H_0^1(\Omega)$. Therefore we can extract a subsequence, denoted in the same way, such that $z_k \rightharpoonup z$ weakly in $H_0^1(\Omega)$ and strongly in $L^{\frac{2p}{p-2}}(\Omega)$. Thanks to this convergence and to (5.10), we get the strong convergences in $L^2(\Omega)$:

$$\frac{\partial a}{\partial y}(x,\bar{y}+\theta_k(y_k-\bar{y}))z_k\nabla\bar{y} \to \frac{\partial a}{\partial y}(x,\bar{y})z\nabla\bar{y} \text{ and } \frac{\partial f}{\partial y}(x,\bar{y}+\nu_k(y_k-\bar{y}))z_k \to \frac{\partial f}{\partial y}(x,\bar{y})z.$$

Therefore we can pass to the limit in (5.13) and deduce

$$-\text{div}\left[a(x,\bar{y})\nabla z + \frac{\partial a}{\partial y}(x,\bar{y})z\nabla\bar{y}\right] + \frac{\partial f}{\partial y}(x,\bar{y})z = v. \tag{5.14}$$

Moreover by using (5.13), (5.14), and the uniform convergence $y_k \to \bar{y}$ it is easy to prove that

$$\int_\Omega a(x,\bar{y})|\nabla z_k|^2\,dx \to \int_\Omega a(x,\bar{y})|\nabla z|^2\,dx.$$

This fact, along with the weak convergence of $\{z_k\}_{k=1}^\infty$ in $H_0^1(\Omega)$, implies the strong convergence $z_k \to z$ in $H_0^1(\Omega)$.

Step 2: $v \in C_{\bar{u}}$. Since $\alpha(x) \leq u_k(x) \leq \beta(x)$ a.e., we have that $v_k(x) \geq 0$ if $\bar{u}(x) = \alpha(x)$ and $v_k(x) \leq 0$ if $\bar{u}(x) = \beta(x)$ a.e. Since the set of functions satisfying these sign conditions is convex and closed in $L^2(\Omega)$, then it is weakly closed, and therefore the weak limit v of $\{v_k\}_{k=1}^\infty$ satisfies the sign condition too. It remains to prove that $v(x) = 0$ for a.a. x such that $\bar{d}(x) \neq 0$. From (5.9), by using the mean value theorem we obtain

$$\begin{aligned}\frac{\rho_k}{k} = \frac{1}{k\rho_k}\|u_k - \bar{u}\|^2_{L^2(\Omega)} > \frac{J(u_k) - J(\bar{u})}{\rho_k} \\ = \int_\Omega \frac{\partial L}{\partial y}(x, \bar{y} + \theta_k(y_k - \bar{y}), \bar{u} + \theta_k(u_k - \bar{u}))z_k\,dx \\ + \int_\Omega \frac{\partial L}{\partial u}(x, \bar{y} + \theta_k(y_k - \bar{y}), \bar{u} + \theta_k(u_k - \bar{u}))v_k\,dx.\end{aligned}$$

Taking limits in both sides of the inequality, using (3.4), (5.14), the already proved convergence $z_k \to z$ in $H^1_0(\Omega)$, and integrating by parts, we get

$$\begin{aligned}0 \geq \int_\Omega \left\{\frac{\partial L}{\partial y}(x, \bar{y}, \bar{u})z + \frac{\partial L}{\partial u}(x, \bar{y}, \bar{u})v\right\} dx \\ = \int_\Omega \left\{\bar{\varphi} + \frac{\partial L}{\partial u}(x, \bar{y}, \bar{u})\right\} v\,dx = \int_\Omega \bar{d}(x)v(x)\,dx = \int_\Omega |\bar{d}(x)||v(x)|\,dx,\end{aligned}$$

the last equality being a consequence of proved signs for v and (3.6). The previous inequality implies that $|\bar{d}(x)v(x)| = 0$ holds a.e., and hence $v(x) = 0$ if $\bar{d}(x) \neq 0$, as we wanted to prove.

Step 3*:* $v = 0$. The next step consists of proving that v does not satisfy the first condition of (5.7). This will lead to the identity $v = 0$. By using (5.9), the definition of $\mathcal{L}$, and the fact that $(\bar{y}, \bar{u})$ and (y_k, u_k) satisfy the state equation, we get

$$\begin{aligned}\mathcal{L}(u_k, y_k, \bar{\varphi}) = \mathcal{J}(y_k, u_k) < \mathcal{J}(\bar{y}, \bar{u}) + \frac{1}{k}\|u_k - \bar{u}\|^2_{L^2(\Omega)} \\ = \mathcal{L}(\bar{u}, \bar{y}, \bar{\varphi}) + \frac{1}{k}\|u_k - \bar{u}\|^2_{L^2(\Omega)}.\end{aligned} \tag{5.15}$$

Performing a Taylor expansion up to the second order, we obtain

$$\begin{aligned}\mathcal{L}(u_k, y_k, \bar{\varphi}) = \mathcal{L}(\bar{u} + \rho_k v_k, \bar{y} + \rho_k z_k, \bar{\varphi}) = \mathcal{L}(\bar{u}, \bar{y}, \bar{\varphi}) + \rho_k D_{(y,u)}\mathcal{L}(\bar{u}, \bar{y}, \bar{\varphi})(z_k, v_k) \\ + \frac{\rho_k^2}{2} D^2_{(y,u)}\mathcal{L}(\bar{u} + \theta_k\rho_k v_k, \bar{y} + \theta_k\rho_k z_k, \bar{\varphi})(z_k, v_k)^2.\end{aligned}$$

This equality, along with (5.15) and (5.9), leads to

$$\rho_k D_{(y,u)}\mathcal{L}(\bar{u}, \bar{y}, \bar{\varphi})(z_k, v_k) + \frac{\rho_k^2}{2} D^2_{(y,u)}\mathcal{L}(w_k, \xi_k, \bar{\varphi})(z_k, v_k)^2 < \frac{1}{k}\|u_k - \bar{u}\|^2_{L^2(\Omega)} \leq \frac{\rho_k^2}{k},$$

where we have put $\xi_k = \bar{y} + \theta_k\rho_k z_k$ and $w_k = \bar{u} + \theta_k\rho_k v_k$. It is obvious that $\xi_k \to \bar{y}$ in $W^{1,p}_0(\Omega)$ and $w_k \to \bar{u}$ in $L^\infty(\Omega)$. Dividing the previous inequality by ρ_k^2 and taking into account the expressions obtained for the derivatives of $\mathcal{L}$, we obtain

$$\begin{aligned}\frac{1}{\rho_k}\int_\Omega \bar{H}_u(x)v_k(x)\,dx + \frac{1}{2}\int_\Omega \left\{H^k_{yy}(x)z_k^2(x) + 2H^k_{yu}(x)z_k(x)v_k(x) + H^k_{uu}(x)v_k^2(x)\right\} dx \\ -\frac{1}{2}\int_\Omega \left\{\frac{\partial a}{\partial y}(x, \xi_k)z_k\nabla z_k + \frac{\partial^2 a}{\partial y^2}(x, \xi_k)z_k^2\nabla\xi_k\right\}\nabla\bar{\varphi}\,dx < \frac{1}{k},\end{aligned} \tag{5.16}$$

where

$$H^k_{yy}(x) = H_{yy}(x, \xi_k(x), w_k(x), \bar{\varphi}(x)),$$

with analogous definitions for H^k_{uu} and H^k_{yu}. It is easy to check that

$$\left\{\begin{array}{l} (H^k_{yy}(x), H^k_{yu}(x), H^k_{uu}(x)) \to (\bar H_{yy}(x), \bar H_{yu}(x), \bar H_{uu}(x)) \\ |H^k_{yy}(x)| + |H^k_{yu}(x)| + |H^k_{uu}(x)| \le C \end{array}\right. \qquad \text{for a.e. } x \in \Omega$$

for some constant $C < \infty$. We also have the following convergence properties:

$$\left\{\begin{array}{l} \dfrac{\partial^j a}{\partial y^j}(x, \xi_k) z_k \nabla\bar\varphi \to \dfrac{\partial^j a}{\partial y^j}(x, \bar y) z \nabla\bar\varphi,\ j = 1, 2, \\ \nabla z_k \longrightarrow \nabla z \ \text{ and } \ z_k \nabla\xi_k \longrightarrow z \nabla\bar y. \end{array}\right. \qquad \text{in } L^2(\Omega)^n.$$

Using these properties we can pass to the limit in (5.16) as follows:

$$\begin{aligned} \limsup_{k\to\infty} & \left\{\frac{1}{\rho_k}\int_\Omega \bar H_u(x) v_k(x)\,dx + \frac{1}{2}\int_\Omega H^k_{uu}(x) v_k^2(x)\,dx\right\} \\ & + \frac{1}{2}\int_\Omega [\bar H_{yy}(x) z^2(x) + 2\bar H_{yu}(x) z(x) v(x)]\,dx \\ & - \frac{1}{2}\int_\Omega \left\{\frac{\partial a}{\partial y}(x, \bar y) z \nabla z + \frac{\partial^2 a}{\partial y^2}(x, \bar y) z^2 \nabla \bar y\right\} \nabla\bar\varphi\,dx \le 0. \end{aligned} \tag{5.17}$$

The rest of the proof is devoted to verifying that the above upper limit is bounded from below by $\frac{1}{2}\int_\Omega \bar H_{uu} v_k^2\,dx$. If this is proved, then from (5.17) and (5.5) we deduce that $J''(\bar u)v^2 = D^2_{(y,u)}\mathcal{L}(\bar u, \bar y, \bar\varphi)(z, v)^2 \le 0$. According to (5.7) this is possible only if $v = 0$. The proof of the mentioned lower estimate is quite technical, which makes an important difference with respect to the finite dimension. In our framework the difficulty is due to the fact that we only have a weak convergence $v_k \rightharpoonup v$. To overcome this difficulty we use a convexity argument. In order to achieve this goal the essential tool is the second condition of (5.7).

From (A4) and (A6) we get

$$\|\bar H_{uu} - H^k_{uu}\|_{L^\infty(\Omega)} \le C\left\{\|\bar y - y_k\|_{L^\infty(\Omega)} + \|\bar u - u_k\|_{L^\infty(\Omega)}\right\} \to 0.$$

Using this property, $\|v_k\|_{L^2(\Omega)} = 1$, and the identity $\bar H_u(x) v_k(x) = |\bar H_u(x)||v_k(x)|$, we obtain

$$\begin{aligned} \limsup_{k\to\infty} & \left\{\frac{1}{\rho_k}\int_\Omega \bar H_u(x) v_k(x)\,dx + \frac{1}{2}\int_\Omega H^k_{uu}(x) v_k^2(x)\,dx\right\} \\ = \limsup_{k\to\infty} & \left\{\frac{1}{\rho_k}\int_\Omega |\bar H_u(x)||v_k(x)|\,dx + \frac{1}{2}\int_\Omega \bar H_{uu}(x) v_k^2(x)\,dx\right\} \\ \ge \limsup_{k\to\infty} & \left\{\frac{1}{\rho_k}\int_{\{|\bar H_u(x)| > \tau\}} \left[|\bar H_u(x)||v_k(x)| + \frac{1}{2}\bar H_{uu}(x) v_k^2(x)\right] dx \right. \\ & \left. + \frac{1}{2}\int_{\{|\bar H_u(x)| \le \tau\}} \bar H_{uu}(x) v_k^2(x)\,dx\right\}, \end{aligned} \tag{5.18}$$

where τ is given by (5.7).

Remembering that $\rho_k\|v_k\|_{L^\infty(\Omega)} = \|u_k - \bar{u}\|_{L^\infty(\Omega)} < 1/k$, we deduce the existence of an integer $k_0 > 0$ such that

$$\frac{\|\bar{H}_{uu}\|_{L^\infty(\Omega)}\rho_k\|v_k\|_{L^\infty(\Omega)}}{\tau} < \frac{\|\bar{H}_{uu}\|_{L^\infty(\Omega)}}{k\tau} < 1 \quad \forall k \geq k_0,$$

and therefore

$$\frac{\tau}{\rho_k}|v_k(x)| \geq \|\bar{H}_{uu}\|_{L^\infty(\Omega)} v_k^2(x) \quad \text{for a.e. } x \in \Omega \quad \forall k \geq k_0.$$

Then we have, with the help of the second condition of (5.7),

$$\begin{aligned} &\limsup_{k\to\infty}\left\{\frac{1}{\rho_k}\int_{\{|\bar{H}_u|>\tau\}}\left[|\bar{H}_u||v_k| + \frac{1}{2}\bar{H}_{uu}v_k^2\right]dx + \frac{1}{2}\int_{\{|\bar{H}_u|\leq\tau\}}\bar{H}_{uu}v_k^2\,dx\right\} \\ &\geq \limsup_{k\to\infty}\left\{\int_{\{|\bar{H}_u|>\tau\}}\left[\|\bar{H}_{uu}\|_{L^\infty(\Omega)} + \frac{1}{2}\bar{H}_{uu}\right]v_k^2\,dx + \frac{1}{2}\int_{\{|\bar{H}_u|\leq\tau\}}\bar{H}_{uu}v_k^2\,dx\right\} \\ &\geq \int_{\{|\bar{H}_u|>\tau\}}\left[\|\bar{H}_{uu}\|_{L^\infty(\Omega)} + \frac{1}{2}\bar{H}_{uu}\right]v^2\,dx \\ &+\frac{1}{2}\int_{\{|\bar{H}_u|\leq\tau\}}\bar{H}_{uu}v^2\,dx \geq \frac{1}{2}\int_\Omega \bar{H}_{uu}v^2\,dx. \end{aligned} \tag{5.19}$$

Combining (5.18) and (5.19) we get the sought-after lower estimate.

Step 4*: Final contradiction.* Using that $\|v_k\|_{L^2(\Omega)} = 1$ along with (5.16), (5.17), (5.18), (5.19), the second condition of (5.7), and the fact that $v = 0$, we deduce

$$\begin{aligned} 0 &\geq \limsup_{k\to\infty}\left\{\int_{\{|\bar{H}_u|>\tau\}}\left[\|\bar{H}_{uu}\|_{L^\infty(\Omega)} + \frac{1}{2}\bar{H}_{uu}\right]v_k^2\,dx + \frac{1}{2}\int_{\{|\bar{H}_u|\leq\tau\}}\bar{H}_{uu}v_k^2\,dx\right\} \\ &\geq \limsup_{k\to\infty}\left\{\frac{\|\bar{H}_{uu}\|_{L^\infty(\Omega)}}{2}\int_{\{|\bar{H}_u|>\tau\}}v_k^2\,dx + \frac{\mu}{2}\int_{\{|\bar{H}_u|\leq\tau\}}v_k^2\,dx\right\} \\ &\geq \frac{\min\{\|\bar{H}_{uu}\|_{L^\infty(\Omega)},\mu\}}{2}\limsup_{k\to\infty}\int_\Omega v_k^2\,dx = \frac{\min\{\|\bar{H}_{uu}\|_{L^\infty(\Omega)},\mu\}}{2} > 0, \end{aligned}$$

providing the contradiction that we were looking for. □

We finish this section by formulating a different version of the sufficient second-order optimality conditions which is equivalent to (5.7); see [9, Theorem 4.4] for the proof of this equivalence. This formulation is very useful for numerical purposes.

Theorem 5.4. *Let us assume that $\bar{u}$ is a feasible control for problem* (P)*. We also assume that there exists $\bar{\varphi} \in W_0^{1,p}(\Omega)$ satisfying* (3.4) *and* (3.5)*. Then* (5.7) *holds if and only if there exist $\delta, \sigma > 0$ such that*

$$J''(\bar{u})h^2 \geq \delta\|h\|^2_{L^2(\Omega)} \quad \forall h \in C^\sigma_{\bar{u}}, \tag{5.20}$$

where

$$C^\sigma_{\bar{u}} = \left\{h \in L^2(\Omega) : h(x) = \left\{\begin{array}{lll} \geq 0 & \text{if } \bar{u}(x) = \alpha(x) & \\ \leq 0 & \text{if } \bar{u}(x) = \beta(x) & \text{for a.e. } x \in \Omega \\ = 0 & \text{if } |\bar{d}(x)| > \sigma & \end{array}\right.\right\}.$$

REFERENCES

[1] J.-J. Alibert and J.-P. Raymond, *Boundary control of semilinear elliptic equations with discontinuous leading coefficients and unbounded controls*, Numer. Funct. Anal. Optim., 18 (1997), pp. 235–250.

[2] A. Bejan, *Convection Heat Transfer*, J. Wiley & Sons, New York, 1995.

[3] F. Bonnans, *Second-order analysis for control constrained optimal control problems of semilinear elliptic systems*, Appl. Math. Optim., 38 (1998), pp. 303–325.

[4] F. Bonnans and E. Casas, *Une principe de Pontryagine pour le contrôle des systèmes semilinéaires elliptiques*, J. Differential Equations, 90 (1991), pp. 288–303.

[5] E. Casas, *Control of an elliptic problem with pointwise state constraints*, SIAM J. Control Optim., 24 (1986), pp. 1309–1318.

[6] E. Casas, *Pontryagin's principle for state-constrained boundary control problems of semilinear parabolic equations*, SIAM J. Control Optim., 35 (1997), pp. 1297–1327.

[7] E. Casas and L. A. Fernandez, *Distributed control of systems governed by a general class of quasilinear elliptic equations*, J. Differential Equations, 104 (1993), pp. 20–47.

[8] E. Casas and L. A. Fernández, *Dealing with integral state constraints in boundary control problems of quasilinear elliptic equations*, SIAM J. Control Optim., 33 (1995), pp. 568–589.

[9] E. Casas and M. Mateos, *Second order optimality conditions for semilinear elliptic control problems with finitely many state constraints*, SIAM J. Control Optim., 40 (2002), pp. 1431–1454.

[10] E. Casas and F. Tröltzsch, *Second-order necessary and sufficient optimality conditions for optimization problems and applications to control theory*, SIAM J. Optim., 13 (2002), pp. 406–431.

[11] E. Casas and J. Yong, *Maximum principle for state-constrained optimal control problems governed by quasilinear elliptic equations*, Differential Integral Equations, 8 (1995), pp. 1–18.

[12] E. Casas, L. A. Fernández, and J. Yong, *Optimal control of quasilinear parabolic equations*, Proc. Roy. Soc. Edinburgh Sect. A, 125 (1995), pp. 545–565.

[13] E. Casas, J.-P. Raymond, and H. Zidani, *Pontryagin's principle for local solutions of control problems with mixed control-state constraints*, SIAM J. Control Optim., 39 (2000), pp. 1182–1203.

[14] E. Casas, F. Tröltzsch, and A. Unger, *Second order sufficient optimality conditions for a nonlinear elliptic boundary control problem*, Z. Anal. Anwendungen, 15 (1996), pp. 687–707.

[15] E. Casas, F. Tröltzsch, and A. Unger, *Second order sufficient optimality conditions for some state-constrained control problems of semilinear elliptic equations*, SIAM J. Control Optim., 38 (2000), pp. 1369–1391.

[16] J. Douglas, T. Dupont, and J. Serrin, *Uniqueness and comparison theorems for nonlinear elliptic equations in divergence form*, Arch. Rational Mech. Anal., 42 (1971), pp. 157–168.

[17] J. C. Dunn, *On second order sufficient optimality conditions for structured nonlinear programs in infinite-dimensional function spaces*, in Mathematical Programming with Data Perturbations, A. Fiacco, ed., Marcel Dekker, New York, 1998, pp. 83–107.

[18] M. Giaquinta, *Introduction to Regularity Theory for Nonlinear Elliptic Systems*, Lectures Math. ETH Zurich, Birkhäuser, Basel, 1993.

[19] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer, Berlin, 1998.

[20] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.

[21] I. Hlaváček, M. Křížek, and J. Malý, *On Galerkin approximations of quasilinear nonpotential elliptic problem of a nonmonotone type*, J. Math. Anal. Appl., 184 (1994), pp. 168–189.

[22] O. Klein, P. Philip, J. Sprekels, and K. Wilmański, *Radiation- and convection-driven transient heat transfer during sublimation growth of silicon carbide single crystals*, J. Crystal Growth, 222 (2001), pp. 832–851.

[23] M. Křížek and L. Liu, *On the maximum and comparison principles for a steady-state nonlinear heat conduction problem*, ZAMM Z. Angew. Math. Mech., 83 (2003), pp. 559–563.

[24] J. L. Lions, *Quelques mèthodes des rèsolution des problèmes aux limites non linèaires*, Dunod, Gauthier–Villars, Paris, 1969.

[25] C. Morrey, *Multiple Integrals in the Calculus of Variations*, Springer, New York, 1966.

[26] J. Nečas, *Les méthodes directes en théorie des equations elliptiques*, Academia, Prague, 1967.

[27] J. Nečas, *Introduction to the Theory of Nonlinear Elliptic Equations*, Teubner-Texte Math. 52, BSB Teubner, Leipzig, 1983.

[28] J.-P. Raymond and F. Tröltzsch, *Second order sufficient optimality conditions for nonlinear parabolic control problems with state constraints*, Discrete Contin. Dyn. Syst., 6 (2000), pp. 431–450.

[29] J.-P. Raymond and H. Zidani, *Hamiltonian Pontryagin's principles for control problems governed by semilinear parabolic equations*, Appl. Math. Optim., 39 (1999), pp. 143–177.

[30] G. Stampacchia, *Equations elliptiques du second ordre à coefficients discontinus*, Ann. Inst. Fourier (Grenoble), 15 (1965), pp. 189–258.

[31] F. Tröltzsch, *Optimale Steuerung partieller Differentialgleichungen—Theorie, Verfahren und Anwendungen*, Vieweg, Wiesbaden, 2005.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 2, pp. 719–733

© 2009 Society for Industrial and Applied Mathematics

SYMBOLIC MODELS FOR NONLINEAR CONTROL SYSTEMS: ALTERNATING APPROXIMATE BISIMULATIONS*

GIORDANO POLA† AND PAULO TABUADA†

Abstract. Symbolic models are abstract descriptions of continuous systems in which symbols represent aggregates of continuous states. In the last few years there has been a growing interest in the use of symbolic models as a tool for mitigating complexity in control design. In fact, symbolic models enable the use of well-known algorithms in the context of supervisory control and algorithmic game theory for controller synthesis. Since the 1990s many researchers faced the problem of identifying classes of dynamical and control systems that admit symbolic models. In this paper we make further progress along this research line by focusing on control systems affected by disturbances. Our main contribution is to show that incrementally globally asymptotically stable nonlinear control systems with disturbances admit symbolic models.

Key words. symbolic models, approximate bisimulation, alternating bisimulation, incremental stability, nonlinear systems

AMS subject classifications. 93A30, 68Q85, 93C57, 93C10

DOI. 10.1137/070698580

1. Introduction.

In recent years we have witnessed the development of different symbolic techniques aimed at reducing the complexity of controller synthesis [EFP06]. These techniques are based on the idea that many states can be treated as equivalent, when synthesizing controllers, and can thus be replaced by a symbol. The models resulting from replacing equivalent states by symbols, termed symbolic models, are typically simpler than the original ones, in the sense that they have a lower number of states. In many cases, one can even construct symbolic models with a finite number of states, which is especially useful for controller design. In fact, the use of symbolic models provides a systematic approach (based on well-established techniques of supervisory control [RW87] and algorithmic game theory [AVW03]) to the design of controllers for classes of specifications that traditionally have not been considered in the context of continuous control systems. These include specifications involving regular languages, temporal logic, fairness constraints, etc., which arise in many application domains such as manufacturing systems, flight control systems, heating and ventilation systems, etc. The search for classes of systems admitting symbolic models goes back to the 1990s and was motivated by problems of verification of dynamical and hybrid systems. Alur and Dill showed in [AD90] that timed automata admit symbolic models; this result was then generalized in [ACHH93, NOSY93] to multirate automata and in [HKPV98, PV94] to rectangular automata. More complex continuous dynamics, but simpler discrete dynamics, were considered in [LPS00], where it was shown that o-minimal hybrid systems also admit symbolic models. Symbolic models for control systems were only considered later, and early results were reported in [KASL00, MRO02, FJL02, CW98]. More precise results appeared recently in [TP06, Tab07], where it was shown that discrete-time controllable linear

*Received by the editors July 27, 2007; accepted for publication (in revised form) October 31, 2008; published electronically February 13, 2009. This work has been partially supported by the National Science Foundation CAREER award 0717188.

http://www.siam.org/journals/sicon/48-2/69858.html

†Department of Electrical Engineering, University of California at Los Angeles, Los Angeles, CA 90095-1594 (pola@ee.ucla.edu, tabuada@ee.ucla.edu).

systems admit symbolic models. Most of these results are based on appropriately adapting the notion of bisimulation introduced by Milner [Mil89] and Park [Par81] to the context of continuous and hybrid systems. A different approach emerged recently through the work of [YW00, HMP05, GP07, Tab08], where an approximate version of bisimulation was considered. While (exact) bisimulation requires that observations of the states are identical, the notion of approximate bisimulation relaxes this condition by allowing observations to be close and within a desired precision. This more flexible notion of bisimulation allows the identification of more classes of systems, admitting symbolic models. Indeed, the work in [Tab08] showed that for every asymptotically stabilizable nonlinear control system it is possible to construct a symbolic model, which is based on an approximate notion of simulation (one-sided version of approximate bisimulation). Extensions of the results in [Tab08], from approximate simulation to approximate bisimulation, can be found in [Gir07, PGT08]. In particular [PGT08] showed that, for the class of (incrementally globally) asymptotically stable nonlinear control systems, symbolic models exist which are approximate bisimulation equivalent to control systems, with a precision that can be chosen a priori, as a design parameter. Control systems in the work of [Tab08, Gir07, PGT08] are not affected by exogenous disturbance inputs. However, in many realistic situations, physical processes are characterized by a certain degree of uncertainty which is often modeled by additional disturbance inputs. Building upon [Tab08, Gir07, PGT08], this paper extends the results in [PGT08] to nonlinear control systems influenced by disturbances. The presence of disturbances requires us to replace the notion of approximate bisimulation used in [PGT08] with the notion of alternating approximate bisimulation, inspired by Alur and coworkers' alternating bisimulation [AHKV98]. To the best of the authors' knowledge, alternating approximate bisimulation was never used before in the context of control systems. This novel notion of bisimulation is a critical ingredient of our results since, as illustrated in section 3.2 through a simple example, approximate bisimulation fails to distinguish between the different role played by control inputs and disturbance inputs. Consequently, control strategies synthesized on symbolic models based on notions of bisimulation and approximate bisimulation cannot be transferred to the original models in a way which is robust with respect to disturbance inputs. Alternating approximate bisimulation solves this problem by guaranteeing that control strategies synthesized on symbolic models, based on alternating approximate bisimulations, can be readily transferred to the original model, independently of the particular evolution of the disturbance inputs.

The main contribution of this paper is to show that incrementally globally asymptotically stable control systems affected by exogenous inputs do admit symbolic models.

Since control systems with disturbances can be thought of as arenas for differential games [Isa99], our results also provide an alternative approach to the study of differential games by means of tools developed in computer science (see, e.g., [Zie98, AVW03]).

Similar ideas to the ones of this paper have been recently explored in [PT07] for the class of linear control systems with disturbances. A detailed discussion on relationships between the results of the present paper and the ones in [PT07] can be found in the last section of this paper. Notions of bisimulation for nonlinear control systems with disturbances have also been studied in [vdS04], albeit with a different purpose. While we are interested in the construction of bisimilar models that are *finite*, the work in [vdS04] uses bisimulation to relate *continuous*, and thus infinite, control systems. We defer to the last section of this paper for a discussion on the relationships between the notion of bisimulation employed in [vdS04] and the one used in this paper.

2. Control systems and stability notions.

2.1. Notation. The symbols $\mathbb{Z}$, $\mathbb{N}$, $\mathbb{R}$, $\mathbb{R}^+$, and $\mathbb{R}^+_0$ denote the set of integers, positive integers, reals, positive reals, and nonnegative reals, respectively. The identity map on a set A is denoted by 1_A. Given two sets A and B, if A is a subset of B, we denote by $\imath_A : A \hookrightarrow B$ or simply by $\imath$ the natural inclusion map taking any $a \in A$ to $\imath(a) = a \in B$. Given a function $f : A \to B$ the symbol $f(A)$ denotes the image of A through f, i.e., $f(A) := \{b \in B : \exists a \in A \text{ s.t. } b = f(a)\}$; if $C \subseteq A$, then $f|_C : C \to B$ denotes the restriction of f to C, so that $f|_C(c) = f(c)$ for any $c \in C$. We identify a relation $R \subseteq A \times B$ with the map $R : A \to 2^B$ defined by $b \in R(a)$ if and only if $(a, b) \in R$. Given a relation $R \subseteq A \times B$, R^{-1} denotes the inverse relation of R, i.e., $R^{-1} := \{(b, a) \in B \times A : (a, b) \in R\}$. Given a vector $x \in \mathbb{R}^n$ we denote by x' the transpose of x and by x_i the ith element of x. Furthermore, $\|x\|$ denotes the infinity norm of x. We recall that $\|x\| := \max\{|x_1|, |x_2|, \ldots, |x_n|\}$, where $|x_i|$ is the absolute value of x_i. Given a set $A \subseteq \mathbb{R}^n$, the symbol $\overline{A}$ denotes the topological closure of A. The symbol $\mathcal{B}_\varepsilon(x)$ denotes the closed ball centered at $x \in \mathbb{R}^n$ with radius $\varepsilon \in \mathbb{R}^+_0$, i.e., $\mathcal{B}_\varepsilon(x) = \{y \in \mathbb{R}^n : \|x - y\| \leq \varepsilon\}$. For any $A \subseteq \mathbb{R}^n$ and $\mu \in \mathbb{R}$ define $[A]_\mu := \{a \in A \mid a_i = k_i\mu,\ k_i \in \mathbb{Z},\ i = 1, \ldots, n\}$. By geometrical considerations on the infinity norm, for any $\mu \in \mathbb{R}^+$ and $\lambda \geq \mu/2$ the collection of sets $\{\mathcal{B}_\lambda(q)\}_{q \in [\mathbb{R}^n]_\mu}$ is a covering of $\mathbb{R}^n$, i.e., $\mathbb{R}^n \subseteq \bigcup_{q \in [\mathbb{R}^n]_\mu} \mathcal{B}_\lambda(q)$; conversely for any $\lambda < \mu/2$, $\mathbb{R}^n \nsubseteq \bigcup_{q \in [\mathbb{R}^n]_\mu} \mathcal{B}_\lambda(q)$. A function $f : [a, b] \to \mathbb{R}^n$ is said to be absolutely continuous on $[a, b]$ if for any $\varepsilon \in \mathbb{R}^+$ there exists $\delta \in \mathbb{R}^+$ so that for every $k \in \mathbb{N}$ and for every sequence of points $a \leq a_1 < b_1 < a_2 < b_2 < \cdots < a_k < b_k \leq b$, if $\sum_{i=1\ldots m}(b_i - a_i) < \delta$, then $\sum_{i=1\ldots m} |f(b_i) - f(a_i)| < \varepsilon$. A function $f :]a, b[\to \mathbb{R}^n$ is said to be locally absolutely continuous if the restriction of f to any compact subset of $]a, b[$ is absolutely continuous. Given a measurable function $f : \mathbb{R}^+_0 \to \mathbb{R}^n$, the (essential) supremum of f is denoted by $\|f\|_\infty$. We recall that $\|f\|_\infty := (ess)\sup\{\|f(t)\|, t \geq 0\}$ and f is essentially bounded if $\|f\|_\infty < \infty$. For a given time $\tau \in \mathbb{R}^+$, define f_τ so that $f_\tau(t) = f(t)$ for any $t \in [0, \tau)$, and $f(t) = 0$ elsewhere; f is said to be locally essentially bounded if for any $\tau \in \mathbb{R}^+$, f_τ is essentially bounded. A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be radially unbounded if $f(x) \to \infty$, as $\|x\| \to \infty$. A continuous function $\gamma : \mathbb{R}^+_0 \to \mathbb{R}^+_0$ is said to belong to class $\mathcal{K}$ if it is strictly increasing and $\gamma(0) = 0$; γ is said to belong to class $\mathcal{K}_\infty$ if $\gamma \in \mathcal{K}$ and $\gamma(r) \to \infty$, as $r \to \infty$. A continuous function $\beta : \mathbb{R}^+_0 \times \mathbb{R}^+_0 \to \mathbb{R}^+_0$ is said to belong to class $\mathcal{KL}$ if for each fixed s the map $\beta(r, s)$ belongs to class $\mathcal{K}_\infty$ with respect to r and, for each fixed r, the map $\beta(r, s)$ is decreasing with respect to s and $\beta(r, s) \to 0$, as $s \to \infty$. Given a metric space $(X, \mathbf{d})$, we denote by $\mathbf{d}_h$ the Hausdorff pseudometric induced by $\mathbf{d}$ on 2^X; we recall that for any $X_1, X_2 \subseteq X$, $\mathbf{d}_h(X_1, X_2) := \max\{\vec{\mathbf{d}}_h(X_1, X_2), \vec{\mathbf{d}}_h(X_2, X_1)\}$, where $\vec{\mathbf{d}}_h(X_1, X_2) = \sup_{x_1 \in X_1} \inf_{x_2 \in X_2} \mathbf{d}(x_1, x_2)$ is the directed Hausdorff pseudometric. We recall that the Hausdorff pseudometric $\mathbf{d}_h$ satisfies the following properties for any $X_1, X_2, X_3 \subseteq X$: (i) $X_1 = X_2$ implies $\mathbf{d}_h(X_1, X_2) = 0$; (ii) $\mathbf{d}_h(X_1, X_2) = \mathbf{d}_h(X_2, X_1)$; (iii) $\mathbf{d}_h(X_1, X_3) \leq \mathbf{d}_h(X_1, X_2) + \mathbf{d}_h(X_2, X_3)$.

2.2. Control systems. The class of systems that we consider in this paper is formalized in the following definition.

DEFINITION 2.1. *A control system is a quadruple* $\Sigma = (\mathbb{R}^n, W, \mathcal{W}, f)$, *where*

- $\mathbb{R}^n$ *is the state space;*
- $W = U \times V$ *is the input space, where*
  - $U \subseteq \mathbb{R}^m$ *is the control input space;*
  - $V \subseteq \mathbb{R}^s$ *is the disturbance input space;*

- *$\mathcal{W} = \mathcal{U} \times \mathcal{V}$ is a subset of the set of all measurable and locally essentially bounded functions of time from intervals of the form $]a, b[\subseteq \mathbb{R}$ to W with $a < 0$ and $b > 0$;*
- *$f : \mathbb{R}^n \times W \rightarrow \mathbb{R}^n$ is a continuous map satisfying the following Lipschitz assumption: for every compact set $K \subset \mathbb{R}^n$, there exists a constant $\kappa > 0$ such that*

$$\|f(x, w) - f(y, w)\| \leq \kappa \|x - y\|$$

for all $x, y \in K$ and all $w \in W$.

A locally absolutely continuous curve $\mathbf{x} :]a, b[\rightarrow \mathbb{R}^n$ is said to be a trajectory of Σ if there exists $\mathbf{w} \in \mathcal{W}$ satisfying $\dot{\mathbf{x}}(t) = f(\mathbf{x}(t), \mathbf{w}(t))$ for almost all $t \in\,]a, b[$.

Although we have defined trajectories over open domains, we shall refer to trajectories $\mathbf{x} :[0, \tau] \rightarrow \mathbb{R}^n$ defined on closed domains $[0, \tau]$, $\tau \in \mathbb{R}^+$, with the understanding of the existence of a trajectory $\mathbf{z} :]a, b[\rightarrow \mathbb{R}^n$ such that $\mathbf{x} = \mathbf{z}|_{[0,\tau]}$. We will also write $\mathbf{x}(\tau, x, \mathbf{w})$ to denote the point reached at time $\tau \in]a, b[$ under the input $\mathbf{w}$ from initial condition x; this point is uniquely determined, since the assumptions on f ensure existence and uniqueness of trajectories. Whenever we need to distinguish between a control input value u and a disturbance input value v in $(u, v) \in W$ we slightly abuse notation by writing $f(x, u, v)$ instead of $f(x, (u, v))$. Analogously, whenever we need to distinguish between $\mathbf{u}$ and $\mathbf{v}$ in an input signal $(\mathbf{u}, \mathbf{v}) \in \mathcal{W}$, we write $\mathbf{x}(\tau, x, \mathbf{u}, \mathbf{v})$ instead of $\mathbf{x}(\tau, x, (\mathbf{u}, \mathbf{v}))$.

In some of the subsequent developments we assume that control systems are forward complete. We recall that a control system Σ is *forward complete* if every trajectory is defined on an interval of the form $]a, \infty[$. Sufficient and necessary conditions for a system to be forward complete are given in [AS99]. Simpler, but only sufficient, conditions for forward completeness are also available in the literature; these include linear growth or compact support of the vector field (see, e.g., [LM67]). The results presented in this paper will rely upon the following stability notion.

Definition 2.2 ([Ang02]). *A control system Σ is said to be incrementally globally asymptotically stable (δ-GAS) if it is forward complete and there exists a $\mathcal{KL}$ function β such that for any $t \in \mathbb{R}_0^+$, any $x_1, x_2 \in \mathbb{R}^n$, and any input signal $\mathbf{w} \in \mathcal{W}$ the following condition is satisfied:*

$$\|\mathbf{x}(t, x_1, \mathbf{w}) - \mathbf{x}(t, x_2, \mathbf{w})\| \leq \beta(\|x_1 - x_2\|, t). \tag{2.1}$$

The above definition can be thought of as an incremental version of the classical notion of global asymptotic stability (GAS). Sufficient and necessary conditions for a control system to be δ-GAS, based on dissipation inequalities, can be found in [Ang02].

3. Symbolic models and approximate equivalence notions.

3.1. Alternating transition systems. In this paper we will use the class of alternating transition systems as abstract models of control systems.

Definition 3.1. *An (alternating) transition system is a tuple:*

$$T = (Q, L, \longrightarrow, O, H),$$

consisting of

- *a set of states Q;*

- *A set of labels $L = A \times B$, where*
  - *A is the set of control labels;*
  - *B is the set of disturbance labels;*
- *A transition relation $\longrightarrow \; \subseteq Q \times L \times Q$;*
- *An output set O;*
- *An output function $H : Q \to O$.*

A transition system T is said to be

- *metric if the output set O is equipped with a metric $\mathbf{d} : O \times O \to \mathbb{R}_0^+$;*
- *countable if Q and L are countable sets;*
- *finite if Q and L are finite sets.*

We will follow standard practice and denote by $q \xrightarrow{a,b} p$ a transition from q to p labeled by a and b. Transition systems capture dynamics through the transition relation. For any states $q, p \in Q$, $q \xrightarrow{a,b} p$ simply means that it is possible to evolve or jump from state q to state p under the action labeled by a and b. A transition system can be represented as a graph where circles represent states and arrows represent transitions (see, e.g., Figure 3.1). We will use transition systems as an abstract representation of control systems. There are several different ways in which we can transform control systems into transition systems. We now describe one of these ways which has the property of capturing all the information contained in a control system Σ. Given $\Sigma = (\mathbb{R}^n, U \times V, \mathcal{U} \times \mathcal{V}, f)$ define the transition system

$$T(\Sigma) := (Q, L, \longrightarrow, O, H),$$

where

- $Q = \mathbb{R}^n$;
- $L = A \times B$, where $A = \mathcal{U}$ and $B = \mathcal{V}$;
- $q \xrightarrow{\mathbf{u},\mathbf{v}} p$ if $\mathbf{x}(\tau, q, \mathbf{u}, \mathbf{v}) = p$ for some $\tau \in \mathbb{R}^+$;
- $O = \mathbb{R}^n$;
- $H = 1_{\mathbb{R}^n}$.

In the subsequent developments we will work with a subtransition system of $T(\Sigma)$ obtained by selecting those transitions from $T(\Sigma)$ describing trajectories of duration τ for some chosen $\tau \in \mathbb{R}^+$. This can be seen as a time discretization or sampling process.

Definition 3.2. *Given a control system $\Sigma = (\mathbb{R}^n, U \times V, \mathcal{U} \times \mathcal{V}, f)$ and a parameter $\tau \in \mathbb{R}^+$, define the transition system*

$$T_\tau(\Sigma) := (Q_\tau, L_\tau, \underset{\tau}{\longrightarrow}, O_\tau, H_\tau),$$

where

- *$Q_\tau = \mathbb{R}^n$;*
- *$L_\tau = A_\tau \times B_\tau$, where $A_\tau = \{\mathbf{u} \in \mathcal{U} \,|\, \text{the domain of } \mathbf{u} \text{ is } [0,\tau]\}$ and $B_\tau = \{\mathbf{v} \in \mathcal{V} \,|\, \text{the domain of } \mathbf{v} \text{ is } [0,\tau]\}$;*
- *$q \xrightarrow[\tau]{\mathbf{u},\mathbf{v}} p$ if $\mathbf{x}(\tau, q, \mathbf{u}, \mathbf{v}) = p$;*
- *$O_\tau = \mathbb{R}^n$;*
- *$H_\tau = 1_{\mathbb{R}^n}$.*

Note that $T_\tau(\Sigma)$ is a metric transition system when we regard $O_\tau = \mathbb{R}^n$ as being equipped with the metric $\mathbf{d}(p, q) = \|p - q\|$.

3.2. Alternating and approximate bisimulations. In this section we introduce a notion of approximate equivalence upon which all the results in this paper rely. The following definition has been introduced in [GP07] and in a slightly different formulation in [Tab08].

Definition 3.3. *Given two metric transition systems* $T_1 = (Q_1, L_1, \xrightarrow[1]{}, O, H_1)$ *and* $T_2 = (Q_2, L_2, \xrightarrow[2]{}, O, H_2)$ *with the same output set* O *and metric* $\mathbf{d}$*, and given a precision* $\varepsilon \in \mathbb{R}_0^+$*, a relation* $R \subseteq Q_1 \times Q_2$ *is said to be an* ε*-approximate bisimulation relation between* T_1 *and* T_2 *if for any* $(q_1, q_2) \in R$

(i) $\mathbf{d}(H_1(q_1), H_2(q_2)) \leq \varepsilon$*;*

(ii) $q_1 \xrightarrow[1]{l_1} p_1$ *implies existence of* $q_2 \xrightarrow[2]{l_2} p_2$ *such that* $(p_1, p_2) \in R$*;*

(iii) $q_2 \xrightarrow[2]{l_2} p_2$ *implies existence of* $q_1 \xrightarrow[1]{l_1} p_1$ *such that* $(p_1, p_2) \in R$*.*

Moreover, T_1 *is* ε*-approximately bisimilar to* T_2 *if there exists an* ε*-approximate bisimulation relation* R *between* T_1 *and* T_2 *such that* $R(Q_1) = Q_2$ *and* $R^{-1}(Q_2) = Q_1$*.*

Note that when $\varepsilon = 0$, the notion of ε-approximate bisimulation relation is equivalent to the classical notion of Milner [Mil89] and Park [Par81]. The work in [PGT08] showed existence of symbolic models that are approximately bisimilar to δ-GAS control systems (with no disturbance). However, the notion stated in Definition 3.3 and employed in [PGT08] does not capture the different role of control and disturbance inputs in control systems. The following example shows that approximate bisimulations (in the sense of Definition 3.3) cannot be used for control design of systems affected by disturbances.

Example 3.4. *Consider the control system* $\Sigma = (\mathbb{R}, U \times V, \mathcal{U} \times \mathcal{V}, f)$*, where* $U = [1, 2] \subset \mathbb{R}$*,* $V = [0.4, 1] \subset \mathbb{R}$*,* $\mathcal{U} \times \mathcal{V}$ *is the class of all measurable and locally essentially bounded functions taking values in* $U \times V$*, and* $f : \mathbb{R} \times U \times V \rightarrow \mathbb{R}$ *is defined by* $f(x, u, v) = -2x + uv$*. We work in the compact state space*[1] $X = [0, 2]$*. Consider the transition system* $T = (Q, L, \longrightarrow, O, H)$*, where*

- $Q = \{q_1, q_2, q_3\}$*;*
- $L = \{l_1, l_2, l_3\}$*;*
- $q \xrightarrow{l} p$ *is depicted in Figure* 3.1*;*
- $O = \mathbb{R}$*;*
- $H : O \rightarrow \mathbb{R}$ *is defined by* $H(q_1) = 0$*,* $H(q_2) = 1$*, and* $H(q_3) = 2$*.*

Given the desired precision $\varepsilon = 0.6$ *and* $\tau = 1$*, by using the results in* [PGT08]*, it is possible to show that the relation* $R \subset Q_\tau \times Q$ *defined by*

$$R = R_1 \times \{q_1\} \cup R_2 \times \{q_2\} \cup R_3 \times \{q_3\}, \tag{3.1}$$

where $R_1 = [0, 0.6]$*,* $R_2 = [0.4, 1.6]$*, and* $R_3 = [1.4, 2]$ *is a* 0.6*-approximate bisimulation relation between* $T_\tau(\Sigma)$ *and* T*. Furthermore, since* $R(Q_\tau) = Q$ *and* $R^{-1}(Q) = Q_\tau$*, transition systems* $T_\tau(\Sigma)$ *and* T *are* 0.6*-approximately bisimilar.*[2] *Suppose now that the goal is to find a control strategy on* T *such that, starting from state* q_1*, it is possible to reach the set* $\{q_2, q_3\}$ *in one step. By Figure* 3.1*,* $q_1 \xrightarrow{l_2} q_2$ *and* $q_1 \xrightarrow{l_3} q_3$*,*

[1] The set X is invariant for the control system Σ; i.e., $\mathbf{x}(t, x, \mathbf{u}, \mathbf{v}) \in X$ for any $x \in X$, any $(\mathbf{u}, \mathbf{v}) \in \mathcal{U} \times \mathcal{V}$, and any time $t \in \mathbb{R}_0^+$.

[2] Transition system T coincides with transition system $T_{\tau,\eta,\mu}(\Sigma)$ as defined in (7) of [PGT08], with $\tau = 1$, $\eta = 1$, and $\mu = 0.01$. Theorem 4.1 of [PGT08] guarantees that T is 0.6-approximately bisimilar to transition system $T_\tau(\Sigma)$ with $\tau = 1$. Notice that condition (8) of [PGT08] boils down, in this case, to $e^{-2\tau}\varepsilon + \mu + \eta/2 \leq \varepsilon$, which is indeed satisfied.

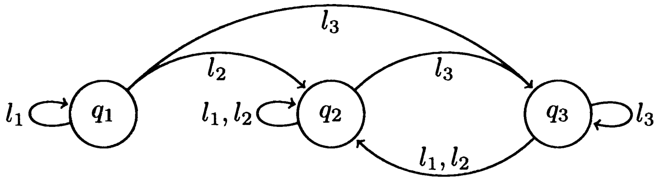


FIG. 3.1. *Transition system T associated with control system Σ of Example* 3.4.

and hence both labels l_2 and l_3 solve that problem. Since $(0,q_1) \in R$, the notion of approximate bisimulation (see condition (iii) *of Definition* 3.3) *guarantees that starting from $0 \in R_1$ there exists a pair of labels $(a_2,b_2),(a_3,b_3) \in A_\tau \times B_\tau$ so that $0 \xrightarrow[\tau]{a_2,b_2} x_2 \in R_2$ and $0 \xrightarrow[\tau]{a_3,b_3} x_3 \in R_3$ in transition system $T_\tau(\Sigma)$. Indeed, by choosing constant curves $(a_2(t),b_2(t)) = (1,1)$ and $(a_3(t),b_3(t)) = (2,1)$, $t \in [0,1]$, we have*

$$0 \xrightarrow[\tau]{a_2,b_2} 0.86 \in R_2, \qquad 0 \xrightarrow[\tau]{a_3,b_3} 1.73 \in R_3. \tag{3.2}$$

However, if the constant disturbance label $b(t) = 0.4$, $t \in [0,1]$, occurs instead of $b_2 = b_3$, we obtain

$$0 \xrightarrow[\tau]{a_2,b} 0.35 \in R_1, \qquad 0 \xrightarrow[\tau]{a_3,b} 0.69 \in R_2,$$

thus showing that the control strategy in (3.2) *does not produce the desired result on the transition system $T_\tau(\Sigma)$. This situation occurs because the notion of bisimulation treats disturbances as cooperative inputs that can be arbitrarily changed to help achieve control objectives. In reality, disturbances need to be treated adversarially. Since $0 \xrightarrow[\tau]{a_3,b} 0.69 \in R_2$ and the set X is invariant for Σ, it is easy to see that for any $\hat{b} \in B_\tau$, $0 \xrightarrow[\tau]{a_3,\hat{b}} x$ with $x \geq 0.69$ and hence $x \in R_2 \cup R_3$. Therefore, control label a_3 guarantees that state $0 \in R_1$ reaches $R_2 \cup R_3$, robustly with respect to the disturbance labels' action, whereas control label a_2 does not. We stress that this different feature of control labels a_2 and a_3 is not captured by the notion of approximate bisimulation in Definition* 3.3.

The above example motivates us to propose the following definition that combines the notions of [GP07] and [Tab08] with the notion of alternating bisimulation, introduced by Alur and coworkers in [AHKV98].

DEFINITION 3.5. *Given two metric transition system $T_1 = (Q_1, A_1 \times B_1, \xrightarrow[1]{}, O, H_1)$ and $T_2 = (Q_2, A_2 \times B_2, \xrightarrow[2]{}, O, H_2)$ with the same observation set O and the same metric $\mathbf{d}$ and given a precision $\varepsilon \in \mathbb{R}_0^+$, a relation $R \subseteq Q_1 \times Q_2$ is said to be an alternating ε-approximate (AεA) bisimulation relation between T_1 and T_2 if for any $(q_1,q_2) \in R$*

(i) *$\mathbf{d}(H_1(q_1), H_2(q_2)) \leq \varepsilon$;*

(ii) *$\forall a_1 \in A_1\ \exists a_2 \in A_2\ \forall b_2 \in B_2\ \exists b_1 \in B_1$ such that $q_1 \xrightarrow[1]{a_1,b_1} p_1$ and $q_2 \xrightarrow[2]{a_2,b_2} p_2$ with $(p_1,p_2) \in R$;*

(iii) *$\forall a_2 \in A_2\ \exists a_1 \in A_1\ \forall b_1 \in B_1\ \exists b_2 \in B_2$ such that $q_1 \xrightarrow[1]{a_1,b_1} p_1$ and $q_2 \xrightarrow[2]{a_2,b_2} p_2$ with $(p_1,p_2) \in R$.*

Moreover, T_1 is said to be $A\varepsilon A$ bisimilar to T_2 if there exists an $A\varepsilon A$ bisimulation relation R between T_1 and T_2 such that $R(Q_1) = Q_2$ and $R^{-1}(Q_2) = Q_1$.

It is easy to see that Definition 3.3 can be recovered as a special case of Definition 3.5, when the cardinality of each of the sets B_1 and B_2 in transition systems T_1 and T_2 is one. Moreover, when $\varepsilon = 0$, the notion of bisimulation in Definition 3.5 coincides with the 2-player version of the definition proposed in [AHKV98].

Definition 3.5 captures the different role played by control and disturbance labels in the transition systems involved, whereas Definition 3.3 does not. In fact, by [AHKV98] it is possible to show that $A\varepsilon A$ bisimulation relations *preserve control strategies* (see Lemma 1 in [AHKV98]) and hence prevent phenomena illustrated in Example 3.4.

4. Existence of symbolic models.

In this section we present the main result of this paper:

Theorem 4.1. *Consider a control system $\Sigma = (\mathbb{R}^n, U \times V, \mathcal{U} \times \mathcal{V}, f)$. If Σ is δ-GAS and $U \times V$ is compact, then for any desired precision $\varepsilon \in \mathbb{R}^+$ there exist $\tau \in \mathbb{R}^+$ and a countable transition system T that is $A\varepsilon A$ bisimilar to $T_\tau(\Sigma)$.*

The above result is important because it shows the existence of symbolic models for nonlinear control systems *in the presence of disturbances*, and therefore it provides a first step toward the construction of symbolic models with guaranteed approximation properties. Theorem 4.1 relies upon the δ-GAS assumption on the control system considered. This condition is not far from also being necessary. Indeed a counterexample can be found in [PGT08], which shows that unstable (autonomous) control systems do not admit, in general, AεA bisimilar countable symbolic models.[3] The last part of this section will be devoted to the proof of Theorem 4.1 which is based on three steps:

1. We first associate a suitable transition system $T_{\tau,\eta,\mu}(\Sigma)$ to a control system $\Sigma = (\mathbb{R}, U \times V, \mathcal{U} \times \mathcal{V}, f)$ (Definition 4.3).
2. We then prove, under a compactness assumption on $U \times V$, that transition system $T_{\tau,\eta,\mu}(\Sigma)$ is countable (Corollary 4.5).
3. We finally prove, under the δ-GAS assumption on Σ, that $T_{\tau,\eta,\mu}(\Sigma)$ is $A\varepsilon A$ bisimilar to $T_\tau(\Sigma)$ (Theorem 4.6).

Step 1. Given a control system Σ, any $\tau \in \mathbb{R}^+$, $\eta \in \mathbb{R}^+$, and $\mu \in \mathbb{R}^+$, we will define the transition system

$$T_{\tau,\eta,\mu}(\Sigma) := (Q, L, \longrightarrow, O, H). \tag{4.1}$$

Parameters τ, η, and μ in transition system $T_{\tau,\eta,\mu}(\Sigma)$ can be thought of, respectively, as a sampling time, a state space, and an input space quantization. In order to define $T_{\tau,\eta,\mu}(\Sigma)$ we will extract a countable set of states Q from Q_τ and a countable set of labels L from L_τ in such a way that the resulting $T_{\tau,\eta,\mu}(\Sigma)$ is countable and $A\varepsilon A$ bisimilar to $T_\tau(\Sigma)$.

From now on, we denote by $\mathbf{d}_h$ the Hausdorff pseudometric induced by the metric $\mathbf{d}$ of the observation space O_τ of $T_\tau(\Sigma)$. Furthermore, since the output function H_τ of $T_\tau(\Sigma)$ is the identity function, we write $\mathbf{d}(x,y) = \|x - y\|$ instead of $\mathbf{d}(x,y) = \|H_\tau(x) - H_\tau(y)\|$. We start by showing that any subset of $\mathbb{R}^n$ can be arbitrarily well approximated by a subset of the lattice $[\mathbb{R}^n]_\eta$, where η is the precision that we require on the approximation.

[3] The notions of bisimulation employed in this paper and in [PGT08] do coincide for autonomous control systems, i.e., control systems where the cardinality of each of the input sets U and V is one.

LEMMA 4.2. *For any set $X \subseteq \mathbb{R}^n$ and any precision $\eta \in \mathbb{R}^+$ there exists $P \subseteq [\mathbb{R}^n]_\eta$ such that $\mathbf{d}_h(P, X) \leq \eta/2$.*

Proof. By geometrical considerations on the infinity norm, $X \subseteq \bigcup_{p\in[\mathbb{R}^n]_\eta} \mathcal{B}_{\eta/2}(p)$ and therefore for any $x \in X$ there exists $p \in [\mathbb{R}^n]_\eta$ such that $\mathbf{d}(x,p) = \|x - p\| \leq \eta/2$. Denote by $\vartheta : X \to [\mathbb{R}^n]_\eta$ a function that associates to any $x \in X$ a vector $p \in [\mathbb{R}^n]_\eta$ so that $\mathbf{d}(x,p) = \|x-p\| \leq \eta/2$ and set $P = \vartheta(X)$. Notice that, by construction, for any $p \in P$ there exists $x \in X$ such that $\mathbf{d}(x,p) = \|x - p\| \leq \eta/2$ (choose x such that $p = \vartheta(x)$). Then by definition of $\mathbf{d}_h$, the statement holds. □

By the above result, for any given precision $\eta \in \mathbb{R}^+$ we can approximate the state space $Q_\tau = \mathbb{R}^n$ of $T_\tau(\Sigma)$ by means of the countable set $Q := [\mathbb{R}^n]_\eta$. This choice for Q guarantees that for any $x \in Q_\tau$ there exists $q \in Q$ so that $\|x - q\| \leq \eta/2$.

The approximation of the set of labels L_τ of $T_\tau(\Sigma)$ is more involved and requires the notion of reachable set. We recall that given a forward complete control system $\Sigma = (\mathbb{R}^n, U \times V, \mathcal{U} \times \mathcal{V}, f)$, any $\tau \in \mathbb{R}^+$, and $x \in \mathbb{R}^n$, the reachable set of $T_\tau(\Sigma)$ with initial condition $x \in Q_\tau$ is the set $\mathcal{R}(\tau, x)$ of endpoints $\mathbf{x}(\tau, x, \mathbf{a}, \mathbf{b})$ for any $\mathbf{a} \in A_\tau$ and $\mathbf{b} \in B_\tau$ or, equivalently,

$$\mathcal{R}(\tau, x) := \left\{ y \in Q_\tau : x \xrightarrow[\tau]{\mathbf{a},\mathbf{b}} y,\ \mathbf{a} \in A_\tau,\ \mathbf{b} \in B_\tau \right\}. \tag{4.2}$$

Moreover, the reachable set of $T_\tau(\Sigma)$ with initial condition $x \in Q_\tau$ and control label $\mathbf{a} \in A_\tau$ is the set $\mathcal{R}(\tau, x, \mathbf{a})$ of endpoints $\mathbf{x}(\tau, x, \mathbf{a}, \mathbf{b})$ for any $\mathbf{b} \in B_\tau$, i.e.,

$$\mathcal{R}(\tau, x, \mathbf{a}) := \left\{ y \in Q_\tau : x \xrightarrow[\tau]{\mathbf{a},\mathbf{b}} y,\ \mathbf{b} \in B_\tau \right\}. \tag{4.3}$$

The reachable sets in (4.2) and (4.3) are well defined because the control system Σ associated with $T_\tau(\Sigma)$ is assumed to be forward complete. Given any desired precision $\mu \in \mathbb{R}^+$, we approximate L_τ by means of the set $L := A \times B$, where

$$A := \bigcup\nolimits_{q\in Q} A^\mu(q), \qquad B := \bigcup\nolimits_{q\in Q} \bigcup\nolimits_{\mathbf{a}\in A^\mu(q)} B^\mu(q, \mathbf{a}), \tag{4.4}$$

and $A^\mu(q)$ captures the set of control labels that can be applied at the state $q \in Q$, while $B^\mu(q, \mathbf{a})$ captures the set of disturbance labels that can be applied at the state $q \in Q$ when the chosen control label is $\mathbf{a} \in A^\mu(q)$. The definition of sets A and B in (4.4) is *asymmetric*. This asymmetry follows from the notion of $A\varepsilon A$ bisimulation relation that we use, where control labels must be chosen *robustly* with respect to the action of disturbance labels (see conditions (ii) and (iii) in Definition 3.5). Given any $\tau \in \mathbb{R}^+$, define the following sets:

(4.5)

$$\mathtt{A}_\mu(\tau, q) := \{P \in 2^{[\mathbb{R}^n]_\mu} \mid \exists \mathbf{a} \in A_\tau \text{ s.t. } \mathbf{d}_h(P, \mathcal{R}(\tau, q, \mathbf{a})) \leq \mu/2\},$$
$$\mathtt{B}_\mu(\tau, q, \mathbf{a}) := \{p \in [\mathbb{R}^n]_\mu \mid \exists \mathbf{b} \in B_\tau \text{ s.t. } \mathbf{d}(p, \mathbf{x}(\tau, q, \mathbf{a}, \mathbf{b})) = \|p - \mathbf{x}(\tau, q, \mathbf{a}, \mathbf{b})\| \leq \mu/2\}.$$

Notice that for any $P \in \mathtt{A}_\mu(\tau, q)$ there may exist a (possibly infinite) set of control labels $\mathbf{a} \in A_\tau$ so that $\mathbf{d}_h(P, \mathcal{R}(\tau, q, \mathbf{a})) \leq \mu/2$. Analogously, for any $p \in \mathtt{B}_\mu(\tau, q, \mathbf{a})$ there may exist a (possibly infinite) set of disturbance labels $\mathbf{b} \in B_\tau$ so that $\mathbf{d}(p, \mathbf{x}(\tau, q, \mathbf{a}, \mathbf{b})) = \|p - \mathbf{x}(\tau, q, \mathbf{a}, \mathbf{b})\| \leq \mu/2$. In order to define the sets $A^\mu(q)$ and $B^\mu(q, \mathbf{a})$ in (4.4) we consider for any $P \in \mathtt{A}_\mu(\tau, q)$ only *one* control label $\mathbf{a} \in A_\tau$ and, respectively, for any $p \in \mathtt{B}_\mu(\tau, q, \mathbf{a})$ only *one* disturbance label $\mathbf{b} \in B_\tau$, as *representatives* of all control labels and all disturbance labels associated with the set P and the vector p, respectively. The sets $A^\mu(q)$ and $B^\mu(q, \mathbf{a})$ will be defined as the collections of these representative control and disturbance labels, respectively. The choice of representatives is defined by the functions

$$\psi_\mu^{\tau,q} : \mathtt{A}_\mu(\tau,q) \to A_\tau, \qquad \varphi_\mu^{\tau,q,\mathbf{a}} : \mathtt{B}_\mu(\tau,q,\mathbf{a}) \to B_\tau, \tag{4.6}$$

where

- $\psi_\mu^{\tau,q}$ associates to any $P \in \mathtt{A}_\mu(\tau,q)$ one control label[4] $\mathbf{a} = \psi_\mu^{\tau,q}(P) \in A_\tau$ so that $\mathbf{d}_h(P, \mathcal{R}(\tau,q,\mathbf{a})) \leq \mu/2$,
- $\varphi_\mu^{\tau,q,\mathbf{a}}$ associates to any $p \in \mathtt{B}_\mu(\tau,q,\mathbf{a})$ one disturbance label[4] $\mathbf{b} = \varphi_\mu^{\tau,q,\mathbf{a}}(p) \in B_\tau$ so that $\mathbf{d}(p, \mathbf{x}(\tau,q,\mathbf{a},\mathbf{b})) = \|p - \mathbf{x}(\tau,q,\mathbf{a},\mathbf{b})\| \leq \mu/2$.

By the above definition, functions $\psi_\mu^{\tau,q}$ and $\varphi_\mu^{\tau,q,\mathbf{a}}$ are not unique. The sets $A^\mu(q)$ and $B^\mu(q,\mathbf{a})$, appearing in (4.4), can now be defined by

$$A^\mu(q) := \psi_\mu^{\tau,q}(\mathtt{A}_\mu(\tau,q)), \quad B^\mu(q,\mathbf{a}) := \varphi_\mu^{\tau,q,\mathbf{a}}(\mathtt{B}_\mu(\tau,q,\mathbf{a})). \tag{4.7}$$

Since the control system Σ is assumed to be forward complete, the reachable sets $\mathcal{R}(\tau,q,\mathbf{a})$, appearing in (4.5), are nonempty; hence sets $\mathtt{A}_\mu(\tau,q)$ and $\mathtt{B}_\mu(\tau,q,\mathbf{a})$ in (4.5) are nonempty and therefore sets $A^\mu(q)$ and $B^\mu(q,\mathbf{a})$ in (4.7) are nonempty, as well.

We now have all the ingredients to define the transition system in (4.1).

DEFINITION 4.3. *Given a forward complete control system $\Sigma = (\mathbb{R}, U \times V, \mathcal{U} \times \mathcal{V}, f)$, any $\tau \in \mathbb{R}^+$, $\eta \in \mathbb{R}^+$, and $\mu \in \mathbb{R}^+$, we define the transition system*

$$T_{\tau,\eta,\mu}(\Sigma) := (Q, L, \longrightarrow, O, H), \tag{4.8}$$

where

- $Q = [\mathbb{R}^n]_\eta$;
- $L = A \times B$, *where*

$$A = \bigcup\nolimits_{q\in Q} A^\mu(q), \qquad B = \bigcup\nolimits_{q\in Q} \bigcup\nolimits_{\mathbf{a}\in A^\mu(q)} B^\mu(q,\mathbf{a}),$$

and the sets $A^\mu(q)$ and $B^\mu(q,\mathbf{a})$ are defined in (4.7)*;*
- $q \xrightarrow{\mathbf{a},\mathbf{b}} p$ *if* $\mathbf{a} \in A^\mu(q)$, $\mathbf{b} \in B^\mu(q,\mathbf{a})$, *and* $\|p - \mathbf{x}(\tau,q,\mathbf{a},\mathbf{b})\| \leq \eta/2$*;*
- $O = \mathbb{R}^n$*;*
- $H = \iota : Q \hookrightarrow O$.

Transition system $T_{\tau,\eta,\mu}(\Sigma)$ is metric when we regard $O = \mathbb{R}^n$ as being equipped with the metric $\mathbf{d}(p,q) = \|H(p) - H(q)\| = \|p - q\|$; furthermore, note that the metric employed for $T_{\tau,\eta,\mu}(\Sigma)$ is the same one used in transition system $T_\tau(\Sigma)$.

Step 2. Transition system $T_{\tau,\eta,\mu}(\Sigma)$ is not countable, in general, because the set $\mathtt{A}_\mu(\tau,q)$ of (4.5) (which is involved in the definition of sets of labels A and B) is not so.[5] However, if the reachable sets in (4.2) associated to Σ are bounded, we can guarantee countability of $T_{\tau,\eta,\mu}(\Sigma)$.

PROPOSITION 4.4. *Consider a forward complete control system $\Sigma = (\mathbb{R}^n, U \times V, \mathcal{U} \times \mathcal{V}, f)$ and any $\tau \in \mathbb{R}^+$. Suppose that for any $x \in \mathbb{R}^n$ the reachable set*[6] *$\mathcal{R}(\tau,x)$ is bounded. Then, for any $\eta \in \mathbb{R}^+$ and $\mu \in \mathbb{R}^+$ the corresponding transition system $T_{\tau,\eta,\mu}(\Sigma)$ is countable.*

Proof. Since for any $\eta \in \mathbb{R}^+$ the set of states Q of $T_{\tau,\eta,\mu}(\Sigma)$ is countable, we only need to show that A and B are countable. Given any precision $\mu \in \mathbb{R}^+$, for

[4] These control and disturbance labels exist by the definition of the sets $\mathtt{A}_\mu(\tau,q)$ and $\mathtt{B}_\mu(\tau,q,\mathbf{a})$.

[5] Recall that the power set of a countable set is, in general, not countable (see, e.g., [Sto63]).

[6] Note that sets $\mathcal{R}(\tau,x)$ are well defined because of the forward completeness assumption on the control system.

any $q \in Q$ consider the set $P(q) := \{p \in [\mathbb{R}^n]_\mu \ : \ \exists z \in \mathcal{R}(\tau, q) \text{ s.t. } \|p - z\| \leq \mu/2\}$. The set $\mathcal{R}(\tau, q)$ is bounded and therefore the set $P(q)$ is finite. Since for any $q \in Q$, $\mathtt{A}_\mu(\tau, q) \subseteq 2^{P(q)}$, $\mathtt{A}_\mu(\tau, q)$ is finite, and therefore $A^\mu(q) = \psi_\mu^{\tau,q}(\mathtt{A}_\mu(\tau, q))$ is finite as well. Moreover, since A is the union of finite sets $A^\mu(q)$ with q ranging in the countable set $Q = [\mathbb{R}^n]_\eta$, the set A is countable (see, e.g., [Sto63]). With respect to the set B, since for any state $q \in Q$ and any $\mathbf{a} \in A^\mu(q)$, $\mathtt{B}_\mu(\tau, q, \mathbf{a}) \subseteq [\mathbb{R}^n]_\mu$, the set $\mathtt{B}_\mu(\tau, q, \mathbf{a})$ is countable and then $B^\mu(q, \mathbf{a}) = \varphi_\mu^{\tau,q,\mathbf{a}}(\mathtt{B}_\mu(\tau, q, \mathbf{a}))$ is countable as well. Finally, since B is the union of countable sets $B^\mu(q, \mathbf{a})$ with q ranging in the countable set $Q = [\mathbb{R}^n]_\eta$ and $\mathbf{a}$ ranging in the finite set $A^\mu(q)$, the set B is countable (see, e.g., [Sto63]). □

A direct consequence of the above result is that if the state space of Σ is bounded, which is the case in many realistic situations, the transition system $T_{\tau,\eta,\mu}(\Sigma)$ of (4.8) is finite. The following result gives a checkable condition that guarantees countability of $T_{\tau,\eta,\mu}(\Sigma)$.

COROLLARY 4.5. *Consider a forward complete control system* $\Sigma = (\mathbb{R}^n, U \times V, \mathcal{U} \times \mathcal{V}, f)$ *and suppose that* $U \times V$ *is compact. Then, for any* $\tau \in \mathbb{R}^+$, $\eta \in \mathbb{R}^+$, *and* $\mu \in \mathbb{R}^+$ *the corresponding transition system* $T_{\tau,\eta,\mu}(\Sigma)$ *is countable.*

Proof. By Proposition 5.1 of [LSW96], for any $\tau \in \mathbb{R}^+$ and $x \in Q_\tau$ the reachable set $\mathcal{R}(\tau, x)$ is bounded. Hence, the result follows by applying Proposition 4.4. □

Step 3. We can now give the following result, which relates δ-GAS to existence of (not necessarily countable) symbolic models.

THEOREM 4.6. *Consider a control system* $\Sigma = (\mathbb{R}^n, U \times V, \mathcal{U} \times \mathcal{V}, f)$ *and any desired precision* $\varepsilon \in \mathbb{R}^+$. *If* Σ *is δ-GAS, then for any* $\tau \in \mathbb{R}^+$, $\mu \in \mathbb{R}^+$, *and* $\eta \in \mathbb{R}^+$ *satisfying the condition*

$$\beta(\varepsilon, \tau) + \mu + \eta/2 < \varepsilon, \tag{4.9}$$

the corresponding transition system $T_{\tau,\eta,\mu}(\Sigma)$ *is AεA bisimilar to* $T_\tau(\Sigma)$.

Before giving the proof of this result, we point out that if Σ is δ-GAS, there always exist parameters $\tau \in \mathbb{R}^+$, $\eta \in \mathbb{R}^+$, and $\mu \in \mathbb{R}^+$ satisfying condition (4.9). In fact, if Σ is δ-GAS, then there exists a sufficiently large $\tau \in \mathbb{R}^+$ so that $\beta(\varepsilon, \tau) < \varepsilon$; then by choosing sufficiently small values of μ and η, condition (4.9) is fulfilled.

Proof. Consider the relation $R \subseteq Q_\tau \times Q$ defined by $(x, q) \in R$ if and only if $||x - q|| \leq \varepsilon$. By construction $R^{-1}(Q) = Q_\tau$; by geometrical considerations on the infinity norm, $Q_\tau \subseteq \bigcup_{p \in [\mathbb{R}^n]_\eta} \mathcal{B}_{\eta/2}(p)$, and therefore, since by (4.9) $\eta/2 < \varepsilon$, we have that $R(Q_\tau) = Q$. We now show that R is an $A\varepsilon A$ bisimulation relation between $T_\tau(\Sigma)$ and $T_{\tau,\eta,\mu}(\Sigma)$. Consider any $(x, q) \in R$. Condition (i) in Definition 3.5 is satisfied by the definition of R and of the involved metric transition systems. Let us now show that condition (ii) in Definition 3.5 also holds. Since δ-GAS implies forward completeness, reachable sets defined in (4.3) are well defined, for any $\tau \in \mathbb{R}^+$, $x \in Q_\tau$, and $\mathbf{a} \in A_\tau$. Consider any $\mathbf{a}_1 \in A_\tau$. Given any $\mu \in \mathbb{R}^+$, by Lemma 4.2, there exists $P \subseteq [\mathbb{R}^n]_\mu$ such that

$$\mathbf{d}_h(P, \mathcal{R}(\tau, q, \mathbf{a}_1)) \leq \mu/2. \tag{4.10}$$

By inequality (4.10), $P \in \mathtt{A}_\mu(\tau, q)$ and then let $\mathbf{a}_2$ be given by[7] $\mathbf{a}_2 = \psi_\mu^{\tau,q}(P) \in A^\mu(q)$. By (4.10), the definition of $\psi_\mu^{\tau,q}$, and the properties of $\mathbf{d}_h$, we have

$$\mathbf{d}_h(\mathcal{R}(\tau, q, \mathbf{a}_1), \mathcal{R}(\tau, q, \mathbf{a}_2)) \leq \mathbf{d}_h(P, \mathcal{R}(\tau, q, \mathbf{a}_1)) + \mathbf{d}_h(P, \mathcal{R}(\tau, q, \mathbf{a}_2)) \leq \mu. \tag{4.11}$$

[7] Note that depending on the choice of function $\psi_\mu^{\tau,q}$, which is not unique, $\mathbf{a}_2$ can either coincide or not with $\mathbf{a}_1$.

Consider now any disturbance label[8] $\mathbf{b}_2 \in B^{\mu}(q,\mathbf{a}_2) \subset B_\tau$ and set $z = \mathbf{x}(\tau, q, \mathbf{a}_2, \mathbf{b}_2) \in \mathcal{R}(\tau, q, \mathbf{a}_2)$. By inequality (4.11) and the definition of $\mathbf{d}_h$, there exists $z_1 \in \overline{\mathcal{R}(\tau, q, \mathbf{a}_1)}$ such that

$$\mathbf{d}(z_1, z) = \|z_1 - z\| \leq \mu. \tag{4.12}$$

The vector[9] z_1 can be either in $\mathcal{R}(\tau, q, \mathbf{a}_1)$ or in $\overline{\mathcal{R}(\tau, q, \mathbf{a}_1)} \setminus \mathcal{R}(\tau, q, \mathbf{a}_1)$; in both cases, for any $\sigma \in \mathbb{R}^+$ there exists $z_2 \in \mathcal{R}(\tau, q, \mathbf{a}_1)$ such that

$$\mathbf{d}(z_1, z_2) = \|z_1 - z_2\| \leq \sigma. \tag{4.13}$$

(In particular, if $z_1 \in \mathcal{R}(\tau, q, \mathbf{a}_1)$, one can choose $z_1 = z_2$.) Choose $\mathbf{b}_1 \in B_\tau$ such that $z_2 = \mathbf{x}(\tau, q, \mathbf{a}_1, \mathbf{b}_1)$. (Notice that since $z_2 \in \mathcal{R}(\tau, q, \mathbf{a}_1)$, such $\mathbf{b}_1 \in B_\tau$ does exist.) Consider the transition $x \xrightarrow[\tau]{\mathbf{a}_1, \mathbf{b}_1} y$ in $T_\tau(\Sigma)$. Since $Q_\tau \subseteq \bigcup_{q' \in [\mathbb{R}^n]_\eta} \mathcal{B}_{\eta/2}(q')$, there exists $p \in Q = [\mathbb{R}^n]_\eta$ such that

$$\mathbf{d}(z, p) = \|z - p\| \leq \eta/2. \tag{4.14}$$

Thus $q \xrightarrow{\mathbf{a}_2, \mathbf{b}_2} p$ in $T_{\tau,\eta,\mu}(\Sigma)$. Since Σ is δ-GAS, by (4.13), (4.12), and (4.14) the following chain of inequalities holds:

$$\begin{aligned}
\|y - p\| &= \|y - z_2 + z_2 - z_1 + z_1 - z + z - p\| \\
&\leq \|y - z_2\| + \|z_2 - z_1\| + \|z_1 - z\| + \|z - p\| \\
&\leq \beta(||x - q||, \tau) + \|z_2 - z_1\| + \|z_1 - z\| + \|z - p\| \leq \beta(\varepsilon, \tau) + \sigma + \mu + \eta/2.
\end{aligned}$$

By inequality (4.9), there exists a sufficiently small value of $\sigma \in \mathbb{R}^+$ such that $\beta(\varepsilon, \tau) + \sigma + \mu + \eta/2 \leq \varepsilon$, and hence $(y, p) \in R$ and condition (ii) in Definition 3.5 holds. We now show that condition (iii) is also satisfied. Consider any $\mathbf{a}_2 \in A$; since $A \subset A_\tau$, we can choose $\mathbf{a}_1 = \mathbf{a}_2 \in A_\tau$. Consider any $\mathbf{b}_1 \in B_\tau$ and set $z = \mathbf{x}(\tau, q, \mathbf{a}_1, \mathbf{b}_1)$. Since $Q_\tau \subseteq \bigcup_{q' \in [\mathbb{R}^n]_\mu} \mathcal{B}_{\mu/2}(q')$, there then exists $z_1 \in [\mathbb{R}^n]_\mu$ such that

$$\mathbf{d}(z, z_1) = \|z - z_1\| \leq \mu/2. \tag{4.15}$$

Furthermore $z \in \mathcal{R}(\tau, q, \mathbf{a}_1)$ and hence, it is clear that $z_1 \in \mathsf{B}_\mu(\tau, q, \mathbf{a}_1)$ by definition of $\mathsf{B}_\mu(\tau, q, \mathbf{a}_1)$. Then let $\mathbf{b}_2$ be given by[10] $\mathbf{b}_2 = \varphi_\mu^{\tau,q,\mathbf{a}_1}(z_1) = \varphi_\mu^{\tau,q,\mathbf{a}_2}(z_1) \in B^\mu(q, \mathbf{a}_2)$. By definition of function $\varphi_\mu^{\tau,q,\mathbf{a}_2}$ and by setting $z_2 = \mathbf{x}(\tau, q, \mathbf{a}_2, \mathbf{b}_2)$, it follows that

$$\mathbf{d}(z_1, z_2) = \|z_1 - z_2\| \leq \mu/2. \tag{4.16}$$

Since $Q_\tau \subseteq \bigcup_{q' \in [\mathbb{R}^n]_\eta} \mathcal{B}_{\eta/2}(q')$, there exists $p \in Q = [\mathbb{R}^n]_\eta$ such that

$$\mathbf{d}(z_2, p) = \|z_2 - p\| \leq \eta/2, \tag{4.17}$$

[8] Existence of such disturbance label is guaranteed by the nonemptyness of set $B^\mu(q, \mathbf{a}_2)$.

[9] The reachable set $\mathcal{R}(\tau, q, \mathbf{a}_1)$ is, in general, not closed and therefore inequality (4.11) does not guarantee the existence of $z_1 \in \mathcal{R}(\tau, q, \mathbf{a}_1)$, satisfying inequality (4.12). However, by definition of $\mathbf{d}_h$, the vector z_1 is guaranteed to exist in the topological closure of the reachable set $\mathcal{R}(\tau, q, \mathbf{a}_1)$.

[10] Note that depending on the choice of function $\varphi_\mu^{\tau,q,\mathbf{a}_1}$, which is not unique, $\mathbf{b}_2$ can either coincide or not with $\mathbf{b}_1$.

and therefore $q \xrightarrow{\mathbf{a_2},\mathbf{b}_2} p$ in $T_{\tau,\eta,\mu}(\Sigma)$. Consider now the transition $x \xrightarrow[\tau]{\mathbf{a}_1,\mathbf{b}_1} y$ in $T_\tau(\Sigma)$. Since Σ is δ-GAS, by (4.15), (4.16), (4.17), and (4.9), the following chain of inequalities holds:

$$\begin{aligned}\|y-p\| &= \|y-z+z-z_1+z_1-z_2+z_2-p\| \\ &\leq \|y-z\|+\|z-z_1\|+\|z_1-z_2\|+\|z_2-p\| \\ &\leq \beta(\|x-q\|,\tau)+\|z-z_1\|+\|z_1-z_2\|+\|z_2-p\| \leq \beta(\varepsilon,\tau)+\mu+\eta/2<\varepsilon.\end{aligned}$$

Thus $(y,p)\in R$, which completes the proof. □

Finally, by combining Corollary 4.5 and Theorem 4.6, the proof of Theorem 4.1 holds as a straightforward consequence.

5. Discussion. In this paper we showed existence of symbolic models that are $A\varepsilon A$ bisimilar to δ-GAS nonlinear control systems with disturbances. Moreover, the parameter ε describing the precision can be chosen as small as desired.

The results of this paper generalize the work in [PGT08] to control systems influenced by disturbances (compare Theorem 4.6 with Theorem 4.1 of [PGT08]). While Theorem 4.1 of [PGT08] states existence of symbolic models that are approximately bisimilar (in the sense of Definition 3.3) to δ-GAS control systems, Theorem 4.6 shows existence of symbolic models that are $A\varepsilon A$ bisimilar to δ-GAS control systems influenced by disturbances. As pointed out in section 3.2, the results of [PGT08] cannot directly be applied to the case of control systems with disturbances. Indeed as Example 3.4 shows, the symbolic model in (7) of [PGT08] does not capture the different role played by the control inputs and by the disturbance inputs. As a consequence, control strategies synthesized on the symbolic model of [PGT08] *cannot be transferred* to the original system. This paper also shares similar ideas with [PT07]. The work in [PT07] proposes symbolic models for linear control systems with disturbances. The approximation notion employed in [PT07] is $A\varepsilon A$ simulation (one-sided version of $A\varepsilon A$ bisimulation). The results in this paper extend the ones in [PT07] by enlarging the class of control systems from linear to nonlinear; enlarging the class of control inputs from piecewise constant to measurable and locally essentially bounded; and generalizing results from simulation to bisimulation. Bisimulation theory for nonlinear control systems in the presence of disturbances has also been considered in [vdS04]. While the focus in [vdS04] was the reduction of continuous systems to continuous systems with lower dimension in the state space, the focus of the present paper is the *reduction* of continuous systems to symbolic models. This difference in purpose translates to a different notion of bisimulation. The notion proposed in [vdS04] is exact,[11] while the notion in Definition 3.5 is approximate. Moreover, systems related by a bisimulation relation according to [vdS04] have the same inputs, while in this paper they have necessarily different inputs since one system is continuous and the other is countable or finite. Working with different inputs forced us to introduce one additional level of quantification on inputs leading to the notion in Definition 3.5, which has four quantifiers while the notion of bisimulation in [vdS04] has two. This additional level of quantification is also responsible for the more complex construction of finite models with respect to previous work in [PGT08], where only two quantifiers were used since disturbances were absent.

Future work will concentrate on constructive techniques to obtain the symbolic models whose existence was shown in this paper. A first step in this direction can be

[11] We recall that an exact bisimulation relation is an ε–approximate bisimulation relation with $\varepsilon = 0$.

found in [PT08] where the construction of symbolic models for linear control systems affected by disturbances is discussed.

Acknowledgment. The authors would like to thank Antoine Girard (Université Joseph Fourier, France) for stimulating discussions on the topic of this paper.

REFERENCES

[ACHH93] R. Alur, C. Courcoubetis, T. A. Henzinger, and P. H. Ho, *Hybrid automata: An algorithmic approach to the specification and verification of hybrid systems*, in Hybrid Systems, Lecture Notes in Comput. Sci. 736, Springer, Berlin, 1993, pp. 209–229.

[AD90] R. Alur and D. L. Dill, *Automata for modeling real-time systems*, in Automata, Languages and Programming, Lecture Notes in Comput. Sci. 443, Springer, Berlin, 1990, pp. 322–335.

[AHKV98] R. Alur, T. Henzinger, O. Kupferman, and M. Vardi, *Alternating refinement relations*, in Proceedings of the 8th International Conference on Concurrence Theory, Lecture Notes in Comput. Sci. 1466, Springer, Berlin, 1998, pp. 163–178.

[Ang02] D. Angeli, *A Lyapunov approach to incremental stability properties*, IEEE Trans. Automat. Control, 47 (2002), pp. 410–421.

[AS99] D. Angeli and E. D. Sontag, *Forward completeness, unboundedness observability, and their Lyapunov characterizations*, Systems Control Lett., 38 (1999), pp. 209–217.

[AVW03] A. Arnold, A. Vincent, and I. Walukiewicz, *Games for synthesis of controllers with partial observation*, Theoret. Comput. Sci., 303 (2003), pp. 7–34.

[CW98] P. E. Caines and Y. J. Wei, *Hierarchical hybrid control systems: A lattice-theoretic formulation*, Special Issue on Hybrid Systems, IEEE Trans. Automat. Control, 43 (1998), pp. 501–508.

[EFP06] M. B. Egerstedt, E. Frazzoli, and G. J. Pappas, eds., *Special Issue on Symbolic Methods for Complex Control Systems*, IEEE Trans. Automat. Control 51, (2006).

[FJL02] D. Forstner, M. Jung, and J. Lunze, *A discrete-event model of asynchronous quantised systems*, Automatica J. IFAC, 38 (2002), pp. 1277–1286.

[Gir07] A. Girard, *Approximately bisimilar finite abstractions of stable linear systems*, in Hybrid Systems: Computation and Control, Lecture Notes in Comput. Sci. 4416, A. Bemporad, A. Bicchi, and G. Buttazzo, eds., Springer, Berlin, 2007, pp. 231–244.

[GP07] A. Girard and G. J. Pappas, *Approximation metrics for discrete and continuous systems*, IEEE Trans. Automat. Control, 52 (2007), pp. 782–798.

[HKPV98] T. A. Henzinger, P. W. Kopke, A. Puri, and P. Varaiya, *What's decidable about hybrid automata?*, J. Comput. System Sci., 57 (1998), pp. 94–124.

[HMP05] T. A. Henzinger, R. Majumdar, and V. Prabhu, *Quantifying similarities between timed systems*, in Proceedings of the Third International Conference on Formal Modeling and Analysis of Timed Systems, 2005, Lecture Notes in Comput. Sci. 3829, Springer, Berlin, 2005, pp. 226–241.

[Isa99] R. Isaacs, *Differential Games*, Dover, New York, 1999.

[KASL00] X. D. Koutsoukos, P. J. Antsaklis, J. A. Stiver, and M. D. Lemmon, *Supervisory control of hybrid systems*, in Proceedings of the IEEE, 88 (2000), pp. 1026–1049.

[LM67] E. B. Lee and L. Markus, *Foundations of Optimal Control Theory*, SIAM Series in Applied Mathematics, John Wiley and Sons, New York, 1967.

[LPS00] G. Lafferriere, G. J. Pappas, and S. Sastry, *O-minimal hybrid systems*, Math. Control Signals Systems, 13 (2000), pp. 1–21.

[LSW96] Y. Lin, E. P. Sontag, and Y. Wang, *A smooth converse Lyapunov theorem for robust stability*, SIAM J. Control Optim., 34 (1996), pp. 124–160.

[Mil89] R. Milner, *Communication and Concurrency*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[MRO02] T. Moor, J. Raisch, and S. D. O'Young, *Discrete supervisory control of hybrid systems based on l-complete approximations*, J. Discrete Event Dyn. Syst., 12 (2002), pp. 83–107.

[NOSY93] X. Nicollin, A. Olivero, J. Sifakis, and S. Yovine. *An approach to the description and analysis of hybrid systems*, in Hybrid Systems, Lecture Notes in Comput. Sci. 736, Springer, Berlin, 1993, pp. 149–178.

[Par81] D. M. R. Park, *Concurrency and automata on infinite sequences*, in Proceedings of the 5th GI Conference on Theoretical Computer Science, Lecture Notes in Comput. Sci. 104, Springer, London, 1981, pp. 167–183.

[PGT08] G. Pola, A. Girard, and P. Tabuada, *Approximately bisimilar symbolic models for nonlinear control systems*, Automatica, 44 (2008), pp. 2508–2516.

[PT07] G. Pola and P. Tabuada, *Symbolic models for linear control systems with disturbances*, in Proceedings of the 46th IEEE Conference on Decision and Control, New Orleans, LA, 2007, pp. 4643–4647.

[PT08] G. Pola and P. Tabuada, *Symbolic models for nonlinear control systems affected by disturbances*, in 47th IEEE Conference on Decision and Control, Cancun, Mexico, 2008, pp. 251–256.

[PV94] A. Puri and P. Varaiya, *Decidability of hybrid systems with rectangular differential inclusion*, in Proceedings of the 6th International Conference on Computer Aided Verification (CAV '94), Lecture Notes in Comput. Sci. 818, Springer, London, 1994, pp. 95–104.

[RW87] P. J. Ramadge and W. M. Wonham, *Supervisory control of a class of discrete event systems*, SIAM J. Control Optim., 25 (1987), pp. 206–230.

[Sto63] R. R. Stoll, *Set Theory and Logic*, W. H. Freeman, San Francisco, 1963.

[Tab07] P. Tabuada, *Symbolic models for control systems*, Acta Inform., 43 (2007), pp. 477–500.

[Tab08] P. Tabuada, *An approximate simulation approach to symbolic control*, IEEE Trans. Automat. Control, 53 (2008), pp. 1406–1418.

[TP06] P. Tabuada and G. J. Pappas, *Linear time logic control of discrete-time linear systems*, IEEE Trans. Automat. Control, 51 (2006), pp. 1862–1877.

[vdS04] A. J. van der Schaft, *Equivalence of dynamical systems by bisimulation*, IEEE Trans. Automat. Control, 49 (2004), pp. 2160–2172.

[YW00] M. Ying and M. Wirsing, *Approximate bisimilarity*, in Algebraic Methodology and Software Technology, Lecture Notes in Comput. Sci. 1816, Springer, Berlin, 2000, pp. 309–322.

[Zie98] W. Zielonka, *Infinite games on finitely coloured graphs with applications to automata on infinite trees*, Theoret. Comput. Sci., 200 (1998), pp. 135–183.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 2, pp. 734–755

© 2009 Society for Industrial and Applied Mathematics

STATE-CONSTRAINED OPTIMAL CONTROL OF SEMILINEAR ELLIPTIC EQUATIONS WITH NONLOCAL RADIATION INTERFACE CONDITIONS[*]

C. MEYER[†] AND I. YOUSEPT[‡]

Abstract. We consider a control- and state-constrained optimal control problem governed by a semilinear elliptic equation with nonlocal interface conditions. These conditions occur during the modeling of diffuse-gray conductive-radiative heat transfer. The nonlocal radiation interface condition and the pointwise state constraints represent the particular features of this problem. To deal with the state constraints, continuity of the state is shown, which allows us to derive first-order necessary conditions. Afterwards, we establish second-order sufficient conditions that account for strongly active sets and ensure local optimality in an L^2-neighborhood.

Key words. nonlinear optimal control, nonlocal radiation interface conditions, state constraints, first-order necessary conditions, second-order sufficient conditions

AMS subject classifications. 49K20, 35B37, 35J65, 90C48

DOI. 10.1137/070694788

1. Introduction.

In this paper, an optimal control problem is investigated that arises from the sublimation growth of semiconductor single crystals such as silicon carbide (SiC) or aluminum nitrite (AlN). To be more precise, the physical vapor transport (PVT) method is considered, where polycrystalline powder is placed under a low-pressure inert gas atmosphere at the bottom of a cavity inside a crucible. The crucible is heated up to 2000 until 3000 K by induction. Due to the high temperatures and the low pressure, the powder sublimates and crystallizes at a single-crystalline seed located at the cooled top of the cavity, such that the desired single crystal grows into the reaction chamber (see [13, 19] for more details). Here, we focus on the control of the conductive-radiative heat transfer in the reaction chamber, which is denoted by Ω_g. More precisely, we aim at optimizing the temperature gradient in Ω_g by directly controlling the heat source u in $\Omega_s := \Omega \backslash \Omega_g$, where Ω denotes the domain of the entire crucible including the gas phase. Thus, the objective functional, considered here, reads as follows:

$$\text{(P)} \qquad \text{Minimize } J(u,y) := \frac{1}{2}\int_{\Omega_g} |\nabla y - z|^2 \, dx + \frac{\beta}{2}\int_{\Omega_s} u^2 \, dx,$$

where y denotes the temperature, z is the desired temperature gradient, and β is a given positive real number. Because of the high temperatures, it is essential to account for radiation on the outer boundary $\Gamma_0 := \partial\Omega$ and on the interface $\Gamma_r := \overline{\Omega}_s \cap \overline{\Omega}_g$. Thus, y is given by the solution of the stationary heat equation with radiation interface

[*] Received by the editors June 19, 2007; accepted for publication (in revised form) September 29, 2008; published electronically February 20, 2009. This research was partially supported by DFG Research Center Matheon, project C9: Numerical simulation and control of sublimation growth of semiconductor bulk single crystals.

http://www.siam.org/journals/sicon/48-2/69478.html

[†] Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstr. 39, D-10117 Berlin, Germany (meyer@wias-berlin.de).

[‡] Institut für Mathematik, Technische Universität Berlin, Str. des 17. Juni 136, D-10623 Berlin, Germany (yousept@math.tu-berlin.de). This author acknowledges the support of the DFG Research Center "Mathematics for Key Technologies" (FZT 86), Berlin.

and boundary conditions on Γ_r and Γ_0, respectively, as follows:

$$\text{(SL)} \qquad \begin{cases} -\operatorname{div}(\kappa_s \nabla y) = u & \text{in } \Omega_s, \\ -\operatorname{div}(\kappa_g \nabla y) = 0 & \text{in } \Omega_g, \\ \kappa_g \left(\dfrac{\partial y}{\partial n_r}\right)_g - \kappa_s \left(\dfrac{\partial y}{\partial n_r}\right)_s = q_r & \text{on } \Gamma_r, \\ \kappa_s \dfrac{\partial y}{\partial n_0} + \varepsilon\sigma\, |y|^3 y = \varepsilon\sigma\, y_0^4 & \text{on } \Gamma_0, \end{cases}$$

where n_0 is the outward unit normal on Γ_0, and n_r is the unit normal on Γ_r facing outward with respect to Ω_s. Furthermore, σ represents the Boltzmann radiation constant, ε is the emissivity, and κ_s, κ_g denote the thermal conductivities in Ω_s, Ω_g, respectively. Moreover, q_r denotes the additional radiative heat flux on Γ_r, which is discussed in more detail in section 1.2. In addition to the stationary semilinear heat equation, the optimization is subject to the following pointwise state and control constraints:

$$\text{(1.1)} \qquad \begin{array}{rcccll} u_a(x) & \leq & u(x) & \leq & u_b(x) & \text{a.e. in } \Omega_s, \\ y_a(x) & \leq & y(x) & \leq & y_b(x) & \text{a.e. in } \Omega_g, \\ & & y(x) & \leq & y_{\max}(x) & \text{a.e. in } \Omega_s. \end{array}$$

Here, u_a and u_b reflect the minimum and maximum heating power, respectively. Furthermore, $y|_{\Omega_s}$ has to be bounded by $y_{\max}$ to avoid melting of the solid components of the crucible in Ω_s. Finally, the state constraints in Ω_g are required to ensure sublimation of the polycrystalline powder and crystallization at the seed, respectively.

The pointwise inequality constraints on the state and the nonlocal radiation on Γ_r represent the crucial points of the problem. First of all, pointwise state constraints are known to be theoretically and numerically difficult to handle since the associated Lagrange multipliers are in general only regular Borel measures; see Casas [9, 10], Alibert and Raymond [3], and Bergounioux and Kunisch [6]. Second, the nonlinearity in the state equation in (P) is in general not monotone (see, for instance, [21]) such that standard techniques cannot be applied. The analysis of the purely control-constrained counterpart to (P) is already comparatively comprehensive. Based on the results of Laitinen and Tiihonen [14] for the nonlinear state equation, first-order necessary conditions for this problem are derived by Meyer, Philip, and Tröltzsch in [15]. Moreover, in [18], second-order sufficient conditions are established, incorporating a generalized two-norm discrepancy. However, these results cannot immediately be transferred to problem (P) due to the presence of pointwise state constraints. Therefore, the inclusion of state constraints represents the genuine contribution of this paper and requires us to significantly extend the analysis of the aforementioned references. First, the continuity of the solution to (SL) is shown in section 2 by means of results on maximum elliptic regularity by Gröger [12] and Elschner, Rehberg, and Schmidt [11]. Based on this, a duality argument allows us to discuss the adjoint equation involving measures as inhomogeneity (cf. section 4), which leads to the derivation of first-order conditions in a standard way (see section 5). Finally, in section 6, second-order sufficient conditions are established that account for strongly active sets and guarantee local optimality with L^2-quadratic growth in an L^2-neighborhood; i.e., the two-norm discrepancy can be avoided. The associated analysis follows the lines of a very recent contribution by Casas, De Los Reyes, and Tröltzsch [7].

1.1. General assumptions and notation. We start now by introducing the general assumptions of the problem statement including the notation used throughout this paper. If X is a linear normed function space, then we use the notation $\|\cdot\|_X$ for a standard norm used in X. Moreover, we set $X^2 := X \times X$. The dual space of X is denoted by X^*, and for the associated duality pairing, we write $\langle \, . \, , \, . \, \rangle_{X^*,X}$. If it is obvious in which spaces the respective duality pairing is considered, then the subscript is occasionally neglected. Now, given another linear normed space Y, the space of all bounded linear operators from X to Y is called $\mathcal{B}(X,Y)$. For an arbitrary $A \in \mathcal{B}(X,Y)$, the associated adjoint operator is denoted by $A^* \in \mathcal{B}(Y^*,X^*)$, and for its inverse, if it exists, we write $A^{-*} := (A^*)^{-1}$. If X is continuously embedded in Y, we write $X \hookrightarrow Y$. The trace operators on Γ_r and Γ_0 are denoted by τ_r and τ_0, respectively. Throughout the paper, they are considered with different domains and ranges. For simplicity, the associated operators are always called τ_r and τ_0, and we will mention their respective domains and ranges if it is important. Furthermore, to improve readability, we sometimes neglect the trace operators in arguments of boundary integrals. The function $\mathbf{1} \in L^\infty(\Gamma_0)$ satisfies $\mathbf{1}(x) = 1$ a.e. on Γ_0, while $\mathcal{U}$ denotes the set of admissible controls with respect to the control constraints, i.e., $\mathcal{U} = \{u \in L^2(\Omega_s) \mid u_a(x) \leq u(x) \leq u_b(x) \text{ a.e. in } \Omega_s\}$. Further, a function $u \in L^2(\Omega_s)$ is called feasible for (P) if it satisfies the inequality constraints in (1.1). Finally, by c we denote a generic positive constant which can take different values on different occasions. Now, concerning the data specified in (P), we impose the following assumptions.

Assumption 1.1.

($\mathcal{A}_1$) *The domain $\Omega \subset \mathbb{R}^N$, $N \in \{2,3\}$, is a bounded open domain with a Lipschitz boundary Γ_0. Moreover, $\Omega_g \subset \Omega$ is an open subset of Ω with a boundary $\Gamma_r \subset \Omega$. In two dimensions, Γ_r is assumed to be a closed Lipschitz surface which is piecewise $\mathcal{C}^{1,\delta}$, whereas it is of class C^1 in the three-dimensional case. The subdomain Ω_s is defined by $\Omega_s = \Omega \setminus \overline{\Omega}_g$. The distance of Γ_r to Γ_0 is assumed to be positive.*

($\mathcal{A}_2$) *The desired temperature gradient z is given in $L^2(\Omega_g)^N$, and $\beta > 0$ is a fixed constant.*

($\mathcal{A}_3$) *The fixed function $\kappa \in L^\infty(\Omega)$ in the semilinear equation* (SL) *is defined by*

$$\kappa(x) = \begin{cases} \kappa_s(x) & \text{if } x \in \Omega_s, \\ \kappa_g(x) & \text{if } x \in \Omega_g, \end{cases}$$

where κ_s and κ_g represent the thermal conductivity of solid and gas, respectively. In two dimensions, κ_s and κ_g satisfy $\kappa_s \in L^\infty(\Omega_s)$ and $\kappa_g \in L^\infty(\Omega_g)$, while they are uniformly continuous in the case of $N = 3$. Moreover, κ satisfies $\kappa(x) \geq \kappa_{\min}$ a.e. in Ω, with a fixed $\kappa_{\min} \in \mathbb{R}^+ \setminus \{0\}$.

($\mathcal{A}_4$) *By $\varepsilon \in L^\infty(\Gamma_0 \cup \Gamma_r)$ we denote the emissivity satisfying $0 < \varepsilon_{\min} \leq \varepsilon(x) \leq 1$ a.e. on $\Gamma_r \cup \Gamma_0$. The term σ represents the Boltzmann radiation and is assumed to be a positive real number. The inhomogeneity on the boundary Γ_0 is given by a fixed function $y_0 \in L^\infty(\Gamma_0)$ satisfying $y_0(x) \geq \theta > 0$ a.e. on Γ_0.*

($\mathcal{A}_5$) *The bounds in the state constraints are $y_{\max} \in \mathcal{C}(\overline{\Omega}_s)$ and $y_a, y_b \in \mathcal{C}(\overline{\Omega}_g)$ with $y_b(x) > y_a(x) \geq \theta$ for all $x \in \overline{\Omega}_g$, $y_{\max}(x) \geq \theta$ for all $x \in \overline{\Omega}_s$, and $y_{\max}(x) > y_a(x)$ for all $x \in \Gamma_r$. For the control constraints, we assume $u_a, u_b \in L^2(\Omega)$ with $0 \leq u_a(x) < u_b(x)$ a.e. in Ω_s.*

1.2. Some well-known results. In the following, we recall some significant results regarding the nonlocal radiation on Γ_r as well as the solvability of the state equation. The results have been discussed in detail in [21, 14, 15]. We start with the following definition.

DEFINITION 1.1. *The radiative heat flux q_r on Γ_r is defined by*

$$q_r = (I - K)(I - (1 - \varepsilon)K))^{-1}\varepsilon\sigma|y^3|y := G\sigma|y^3|y, \tag{1.2}$$

where the integral operator K is defined by

$$(Ky)(x) = \int_{\Gamma_r} \omega(x, z)y(z)\, ds_z,$$

with a symmetric kernel ω. In the case of a two-dimensional domain, the kernel $\omega : \Gamma_r \times \Gamma_r \to \mathbb{R}$ is formally given by

$$\omega(x, z) = \Xi(x, z)\frac{[n_r(z) \cdot (x - z)][n_r(x) \cdot (z - x)]}{2|z - x|^3} \quad \forall x, z \in \Gamma_r,$$

and in the case of a three-dimensional domain,

$$\omega(x, z) = \Xi(x, z)\frac{[n_r(z) \cdot (x - z)][n_r(x) \cdot (z - x)]}{\pi|z - x|^4} \quad \forall x, z \in \Gamma_r.$$

Notice that $\Xi : \Gamma_r \times \Gamma_r \to \mathbb{R}$ denotes the visibility factor, which is defined by

$$\Xi(x, z) = \begin{cases} 0 & \text{if } \overline{xz} \cap \Omega_g \neq \emptyset, \\ 1 & \text{if } \overline{xz} \cap \Omega_g = \emptyset, \end{cases}$$

where $\overline{xz}$ denotes the line between x and z. For the properties of ω and K we refer the reader to Tiihonen [21]. The following lemma (see [14, Lemma 8] for the proof) provides some significant properties of the operator G, which will be useful for our analysis.

LEMMA 1.1. *The operator $G := (I - K)(I - (1 - \varepsilon)K)^{-1}\varepsilon$ is linear and bounded form $L^p(\Gamma_r)$ to $L^p(\Gamma_r)$ for all $1 \le p \le \infty$.*

In the following, we briefly discuss the existence of solutions of (SL). To that end, let us introduce the space

$$V := \{v \in H^1(\Omega) \mid \tau_r v \in L^5(\Gamma_r), \tau_0 v \in L^5(\Gamma_0)\}.$$

Moreover, we define the operator associated with the left-hand side of (SL) that is formally obtained by integration of (SL) by parts over the boundaries Γ_r and Γ_0.

DEFINITION 1.2. *The operator $A : V \to V^*$ is given by*

$$\langle A(y), v\rangle := \int_\Omega \kappa\nabla y \cdot \nabla v\, dx + \int_{\Gamma_r} (G\sigma|y|^3 y)v\, ds + \int_{\Gamma_0} \varepsilon\sigma|y|^3 yv\, ds, \quad y, v \in V,$$

with $G : L^{5/4}(\Gamma_r) \to L^{5/4}(\Gamma_r)$.

Notice that, thanks to the definition of V, the operator A is well defined and continuous. Furthermore, $E_s : L^2(\Omega_s) \to V^*$ and $E_0 : L^{5/4}(\Gamma_0) \to V^*$ are defined by

$$\langle E_s\, u, v\rangle := \int_{\Omega_s} uv\, dx, \quad v \in V, \quad \text{and} \quad \langle E_0\, z\rangle := \int_{\Gamma_0} zv\, ds, \quad v \in V. \tag{1.3}$$

Clearly, E_s and E_0 are linear and bounded in their respective spaces.

Definition 1.3. *A function $y \in V$ is said to be a (weak) solution of* (SL) *if y satisfies the following operator equation:*

$$A(y) = E_s u + E_0\, \varepsilon\sigma y_0^4 \quad in \quad V^*. \tag{1.4}$$

To show the existence of solutions according to this definition, the theory of monotone operators is not applicable here, since the operator G is not positive, i.e., $v(x) \geq 0$ a.e. on Γ_r does not in general imply $(Gv)(x) \geq 0$ a.e. on Γ_r; see [14] for details. However, the existence of weak solutions can be verified by Brezis' theorem for pseudomonotone operators; cf. [22]. In fact, Laitinen and Tiihonen showed in [14] that A is pseudomonotone giving in turn the existence of weak solutions of (1.4). The uniqueness then follows from a comparison principle (cf. [14]). Furthermore, Meyer, Philip, and Tröltzsch [15] showed the boundedness of the solution. We summarize these results in the following theorem.

Theorem 1.1. *Let Assumption* 1.1 *be satisfied. Then for each $u \in L^2(\Omega_s)$, there exists a unique solution $y \in V$ to* (SL) *in the sense of Definition* 1.3*. Moreover, the solution is bounded, i.e., $y \in L^\infty(\Omega)$, and satisfies*

$$\|y\|_{L^\infty(\Omega)} + \|y\|_{L^\infty(\Gamma_r \cup \Gamma_0)} \leq c(\Omega)(1 + \|u\|_{L^2(\Omega_s)} + \|y_0\|^4_{L^{16}(\Gamma_0)}) \tag{1.5}$$

with some constant $c(\Omega) > 0$.

2. Continuous solutions. Our goal in the upcoming sections consists of providing the first-order necessary optimality conditions for (P). To accomplish this task, we will utilize the Karush–Kuhn–Tucker (KKT) theory (see section 5 below). Mainly, we follow the lines of [10]. However, to apply this technique, one has to consider the state constraints in a space such that the convex set, defined by these constraints, admits a nonempty interior. Here, we choose the space of continuous functions, denoted by $\mathcal{C}(\overline{\Omega})$. Therefore, it is at first necessary to show the continuity of the solutions to (SL). The subsequent analysis follows a classical bootstrapping argument. Based on Theorem 1.1, one shows that (SL) admits solutions in the space $W^{1,q}(\Omega)$ with $q > N$. Afterwards the continuous embedding $W^{1,q}(\Omega) \hookrightarrow \mathcal{C}(\overline{\Omega})$, $q > N$, implies the desired continuity. We start with a lemma that represents the key point within the proof of continuity.

Lemma 2.1. *There is a positive real number $\hat{q}$ with $N < \hat{q} < 6$ such that the operator $B(f) : W^{1,q}(\Omega) \to W^{1,q'}(\Omega)^*$, $1/q + 1/q' = 1$, defined by*

$$\langle B(f)y, v\rangle := \int_\Omega \kappa \nabla y \cdot \nabla v \, dx + \int_{\Gamma_0} f y v \, ds, \quad y \in W^{1,q}(\Omega), \quad v \in W^{1,q'}(\Omega),$$

is continuously invertible for all $q \in [N, \hat{q}]$ and all nonnegative functions $f \in L^\infty(\Gamma_0)$ that are positive on a set of measure greater than zero.

Proof. In the two-dimensional case, $N = 2$, the assertion is an immediate consequence of a result of Gröger [12, Theorem 1]. In three dimensions, $N = 3$, we apply a result of Elschner, Rehberg, and Schmidt [11]. First, the Lax–Milgram lemma implies that for every functional $g \in H^1(\Omega)^*$, there exists a unique solution $y \in H^1(\Omega)$ of

$$\int_\Omega \kappa \nabla y \cdot \nabla v \, dx + \int_{\Gamma_0} f y v \, ds = g(v) \quad \forall v \in H^1(\Omega). \tag{2.1}$$

Let $g \in W^{1,q'}(\Omega)^*$ be arbitrarily fixed. Since $q \geq 2$, the dual space $W^{1,q'}(\Omega)^*$ is continuously embedded in $H^1(\Omega)^*$. Consequently, there exists a unique solution $y \in H^1(\Omega)$ of (2.1) with $g \in W^{1,q'}(\Omega)^*$ in the right-hand side of (2.1).

Now, consider the following equation:

$$\int_\Omega \kappa\nabla\eta\cdot\nabla v\,dx + \int_\Omega \eta v\,dx = g(v) - \int_{\Gamma_0} fyv\,ds + \int_\Omega yv\,dx \quad \forall\, v\in W^{1,q'}(\Omega). \tag{2.2}$$

Due to $y\in H^1(\Omega)$ and $N=3$, it holds that $y\in L^6(\Omega)$ and $\tau_0\, y\in L^4(\Gamma_0)$. Hence, since $f\in L^\infty(\Gamma_0)$, we have $fy\in L^{4/3}(\Gamma_0)^*$. For this reason, since $q'\in[6/5,3/2]$ and because of the continuity of the trace operator from $W^{1,6/5}(\Omega)$ to $L^{4/3}(\Gamma_0)$ for $N=3$ (see [1]), the right-hand side of (2.2) defines an element $\xi\in W^{1,q'}(\Omega)^*$ with

$$\langle \xi, v\rangle_{W^{1,q'}(\Omega)^*,W^{1,q'}(\Omega)} := g(v) - \int_{\Gamma_0} fyv\,ds + \int_\Omega yv\,dx, \quad v\in W^{1,q'}(\Omega).$$

Therefore, in view of our assumptions on Ω for $N=3$ (cf. Assumption 1.1) and Remark 3.18 in [11], there exists a real number $\hat q>3$ (independent of f,g) such that for all $q\in[3,\hat q]$, (2.2) admits a unique solution $\eta\in W^{1,q}(\Omega)$. Moreover, the solution can be estimated by

$$\begin{aligned}\|\eta\|_{W^{1,q}(\Omega)} &\le c\|\xi\|_{W^{1,q'}(\Omega)^*}\\ &\le c\left(\|g\|_{W^{1,q'}(\Omega)^*} + (1+\|f\|_{L^\infty(\Gamma_0)})\|y\|_{H^1(\Omega)}\right) \le c\,\|g\|_{W^{1,q'}(\Omega)^*},\end{aligned} \tag{2.3}$$

with a constant $c>0$ independent of g. Clearly, due to $H^1(\Omega)\subset W^{1,q'}(\Omega)$, η also solves

$$\int_\Omega \kappa\nabla\eta\cdot\nabla v\,dx + \int_\Omega (\eta-y)v\,dx = g(v) - \int_{\Gamma_0} fyv\,ds \quad \forall\, v\in H^1(\Omega).$$

Subtracting (2.1) from the above equation and inserting $v=y-\eta$ in the resulting equation, we have

$$\min\{\kappa_{\min},1\}\,\|\eta-y\|^2_{H^1(\Omega)} \le \int_\Omega \kappa|\nabla(\eta-y)|^2 dx + \int_\Omega (\eta-y)^2 dx = 0. \tag{2.4}$$

Notice that we have used $(\mathcal{A}_3)$ in Assumption 1.1 for the latter inequality. Obviously, (2.4) implies that $\eta(x)=y(x)$ a.e. in Ω and a.e. on $\Gamma_r\cup\Gamma_0$. Therefore, possibly after a modification on a set of measure zero, we have $y=\eta$ in $W^{1,q}(\Omega)$.

Thus, for $q\in[3,\hat q]$, the operator equation

$$B(f)y = g \quad \text{in } W^{1,q'}(\Omega)^*$$

admits a unique solution in $W^{1,q}(\Omega)$ for every given $g\in W^{1,q'}(\Omega)^*$. Moreover, (2.3) yields the continuity of $B(f)^{-1}: W^{1,q'}(\Omega)^*\to W^{1,q}(\Omega)$. □

For the rest of this paper, let us fix an arbitrary $q\in(N,\hat q)$. Next, let us redefine the notion of weak solutions of (SL).

Definition 2.1. *The operator $A_q : W^{1,q}(\Omega)\to W^{1,q'}(\Omega)^*$ is defined by*

$$\langle A_q(y), v\rangle := \int_\Omega \kappa\nabla y\cdot\nabla v\; dx + \int_{\Gamma_r} (G\sigma|y|^3 y)v\; ds + \int_{\Gamma_0} \varepsilon\sigma|y|^3 yv\; ds$$

with $y\in W^{1,q}(\Omega)$, $v\in W^{1,q'}(\Omega)$, and $G: L^\infty(\Gamma_r)\to L^\infty(\Gamma_r)$. Moreover, similarly to (1.3), *the operators $E_{q,s}: L^2(\Omega_s)\to W^{1,q'}(\Omega)^*$ and $E_{q,0}: L^\infty(\Gamma_0)\to W^{1,q'}(\Omega)^*$ are given by*

$$\langle E_{q,s}\,u, v\rangle := \int_{\Omega_s} uv\,dx, \quad v\in W^{1,q'}(\Omega), \quad \textit{and} \quad \langle E_{q,0}\,z, v\rangle := \int_{\Gamma_0} zv\,ds, \quad v\in W^{1,q'}(\Omega).$$

Then, analogously to Definition 1.3*, a function* $y \in W^{1,q}(\Omega)$ *is said to be a (weak) solution of* (SL) *if it fulfills the operator equation*

$$A_q(y) = E_{q,s}\, u + E_{q,0}\, \varepsilon\sigma y_0^4 \quad \textit{in} \quad W^{1,q'}(\Omega)^*. \tag{2.5}$$

Notice that A_q is well defined since $y \in W^{1,q}(\Omega)$, $q > N$, implies $\tau_r y \in L^\infty(\Gamma_r)$ and $\tau_0 y \in L^\infty(\Gamma_0)$. Moreover, $E_{q,s} : L^2(\Omega_s) \to W^{1,q'}(\Omega)^*$ is continuous because of $W^{1,q'}(\Omega) \hookrightarrow L^{s'}(\Omega)$ with $s' = \frac{N\,q'}{N-q'} = \frac{N\,q}{(N-1)q-q} \geq 2$ for $q \leq 6$ (cf., for instance, [1]).

Theorem 2.1. *For every* $u \in L^2(\Omega_s)$*, there exists a unique weak solution* $y \in W^{1,q}(\Omega)$ *of* (SL) *in the sense of Definition* 2.1*. Moreover, the following estimate holds true:*

$$\|y\|_{W^{1,q}(\Omega)} \leq c\left(1 + \|u\|_{L^2(\Omega_s)} + \|y_0\|^4_{L^\infty(\Gamma_0)} + \|u\|^4_{L^2(\Omega_s)} + \|y_0\|^{16}_{L^\infty(\Gamma_0)}\right) \tag{2.6}$$

with a constant $c > 0$ *independent of* u, y_0.

Proof. As stated above, we apply Lemma 2.1 to the state equation (SL). First, we observe that the solution of (1.4) for an arbitrary $u \in L^2(\Omega_s)$, again denoted by y, solves

$$\begin{aligned} \int_\Omega \kappa \nabla y \cdot \nabla v \, dx + \int_{\Gamma_0} yv \, ds = \int_{\Omega_s} uv \, dx &- \int_{\Gamma_r} \alpha_G(y) v \, ds \\ &+ \int_{\Gamma_0} (\varepsilon\sigma y_0^4 + \alpha_0(y)) v \, ds \quad \forall\, v \in V \end{aligned} \tag{2.7}$$

with $\alpha_0(y) := y - \varepsilon\sigma|y|^3 y$ and $\alpha_G(y) := G\sigma|y|^3 y$. Due to Theorem 1.1, we have $\alpha_0(y) \in L^\infty(\Gamma_0)$ and $\alpha_G(y) \in L^\infty(\Gamma_r)$. Now, let us consider the following equality:

$$\langle B(\mathbf{1})\eta, v\rangle = \int_{\Omega_s} uv \, dx - \int_{\Gamma_r} \alpha_G(y) v \, ds + \int_{\Gamma_0} (\varepsilon\sigma y_0^4 + \alpha_0(y)) v \, ds \quad \forall\, v \in W^{1,q'}(\Omega). \tag{2.8}$$

Lemma 2.1 implies that $B(\mathbf{1})^{-1} \in \mathcal{B}(W^{1,q'}(\Omega)^*, W^{1,q}(\Omega))$. Moreover, the right-hand side in (2.8), denoted by ω_y, defines a functional in $W^{1,q'}(\Omega)^*$, which is demonstrated in the following. As mentioned above, embedding theorems imply $W^{1,q'}(\Omega) \hookrightarrow L^2(\Omega)$ if $q \leq 6$. Moreover, the trace operators τ_r and τ_0 are continuous from $W^{1,q'}(\Omega)$ to $L^{r'}(\Gamma_r)$ and $L^{r'}(\Gamma_0)$, respectively, with $r' = \frac{(N-1)q'}{N-q'} = \frac{(N-1)q}{(N-1)q-N} > 1$ (see again [1]). Hence, Hölder's inequality implies

$$\begin{aligned} \|\omega_y\|_{W^{1,q'}(\Omega)^*} &= \sup_{\|v\|_{W^{1,q'}}=1} \left| \int_{\Omega_s} uv \, dx - \int_{\Gamma_r} \alpha_G(y) v \, ds + \int_{\Gamma_0} (\varepsilon\sigma y_0^4 + \alpha_0(y)) v \, dx \right| \\ &\leq \sup_{\|v\|_{W^{1,q'}}=1} \Big(\|u\|_{L^2(\Omega_s)} \|v\|_{L^2(\Omega)} + \|\alpha_G(y)\|_{L^\infty(\Gamma_r)} \|v\|_{L^1(\Gamma_r)} \\ &\qquad + \big(\|\varepsilon\sigma y_0^4\|_{L^\infty(\Gamma_0)} + \|\alpha_0(y)\|_{L^\infty(\Gamma_0)} \big) \|v\|_{L^1(\Gamma_0)} \Big) \\ &\leq c\big(\|u\|_{L^2(\Omega_s)} + \|G\|_{\mathcal{B}(L^\infty(\Gamma_r))} \|y\|^4_{L^\infty(\Gamma_r)} \\ &\qquad + \|y_0\|^4_{L^\infty(\Gamma_0)} + \|y\|_{L^\infty(\Gamma_0)} + \|y\|^4_{L^\infty(\Gamma_0)} \big), \end{aligned} \tag{2.9}$$

with a constant $c > 0$ independent of u, y_0, and y. Together with (1.5), the latter inequality ensures $\|\omega_y\|_{W^{1,q}(\Omega)^*} < \infty$. Therefore, (2.8) admits a unique solution

$\eta \in W^{1,q}(\Omega)$, satisfying

$$\|\eta\|_{W^{1,q}(\Omega)} \leq \|B(\mathbf{1})^{-1}\|_{\mathcal{B}(W^{1,q'}(\Omega)^*,W^{1,q}(\Omega))}\,\|\omega_y\|_{W^{1,q'}(\Omega)^*}. \tag{2.10}$$

An argument analogous to the proof of Lemma 2.1 implies that

$$\eta = y \text{ in } W^{1,q}(\Omega).$$

For this reason, $y \in W^{1,q}(\Omega)$ is the unique solution of

$$B(\mathbf{1})y = \omega_y \text{ in } W^{1,q'}(\Omega)^*. \tag{2.11}$$

Thanks to the definitions of $B(\mathbf{1})$, $\alpha_0(y)$, and $\alpha_G(y)$, (2.11) is equivalent to

$$\begin{aligned}\int_\Omega \kappa\nabla y\cdot\nabla v\;dx + \int_{\Gamma_r}(G\sigma|y|^3y)v\;ds + \int_{\Gamma_0}\varepsilon\sigma|y|^3yv\;ds\\ = \int_{\Omega_s} uv\,dx + \int_{\Gamma_0}\varepsilon\sigma y_0^4 v\,ds \quad \forall\, v\in W^{1,q'}(\Omega).\end{aligned}$$

Hence, for every $u \in L^2(\Omega_s)$, (2.5) admits a unique solution $y \in W^{1,q}(\Omega)$. Finally, (2.6) follows from (2.10) together with (2.9) and (1.5). □

Corollary 2.1. *Thanks to* $W^{1,q}(\Omega) \hookrightarrow \mathcal{C}(\overline{\Omega})$*, the solution of* (SL) *is continuous.*

Remark 2.1. Note that additional mapping properties of G are needed to show continuity of the solution with a similar bootstrapping technique and an increased differentiability index, e.g., $H^2(\Omega) \hookrightarrow \mathcal{C}(\overline{\Omega})$. Then, the right-hand side on the interface boundary in (2.8) must be an element of $H^{1/2}(\Gamma_r)$, and therefore one would need $G : H^{1/2}(\Gamma_r) \to H^{1/2}(\Gamma_r)$ which is in general not evident. It might be possible to use results of Agmon, Douglas, and Nirenberg [2], together with density arguments, and a similar bootstrapping technique for the nonlocal radiation, as carried out above, to prove Lemma 2.1. However, it is doubtful if the results can be improved in this way (cf. the counterexamples in [11, section 4] and [20]).

Based on Theorem 2.1, we define the control-to-state operator $\mathcal{G} : L^2(\Omega_s) \to W^{1,q}(\Omega)$ associated with (P), i.e., the solution operator for (SL), that assigns to each $u \in L^2(\Omega_s)$ the weak solution $y \in W^{1,q}(\Omega)$. With this setting at hand, the optimal control problem can equivalently be stated as follows:

$$\text{(P)}\qquad \begin{cases}\min\limits_{u\in\mathcal{U}} & f(u) := J(u,\mathcal{G}(u))\\ \text{subject to} & y_a(x) \leq \mathcal{G}(u)(x) \leq y_b(x) \quad \forall\, x\in\overline{\Omega}_g,\\ & \qquad\qquad \mathcal{G}(u)(x) \leq y_{\max}(x) \quad \forall\, x\in\overline{\Omega}_s,\end{cases}$$

Notice that the reduced objective functional $f(u)$ and the state constraints are well defined since $\mathcal{G}(u) \in W^{1,q}(\Omega) \subset H^1(\Omega)\cap\mathcal{C}(\overline{\Omega})$.

3. Differentiability of the control-to-state operator. Next, let us turn to the linearized version of (SL). First, recall a result of Meyer, Philip, and Tröltzsch [15], that is, the following maximum principle.

Lemma 3.1. *Suppose that* $u \in L^2(\Omega_s)$ *satisfies* $u(x) \geq 0$ *a.e. in* Ω_s*, while* $y_0 \in L^\infty(\Gamma_0)$ *fulfills* $y_0(x) \geq \theta > 0$ *a.e. on* Γ_0 *according to Assumption* 1.1*. Then, the weak solution* y *of* (SL) *satisfies* $y(x) \geq \theta > 0$ *a.e. in* Ω *and a.e. on* Γ_r *and* Γ_0*.*

Now, let $\bar{u} \in L^2(\Omega_s)$ with associated state $\bar{y} \in W^{1,q}(\Omega)$. Moreover, we assume for the rest of this section that $\bar{u}(x) \geq 0$ a.e. in Ω_s such that Lemma 3.1 implies

$\bar{y}(x) > 0$ a.e. on Γ_r and Γ_0. Next, we turn to the derivative of the operator A_q, as given in Definition 2.1, at the point $\bar{y}$. We already mentioned that τ_r and τ_0 are continuous from $W^{1,q}(\Omega)$ to $L^\infty(\Gamma_r)$ and $L^\infty(\Gamma_0)$, respectively. Furthermore, the Nemyzki operator $\Phi(y) := |y|^3 y$ is continuously Fréchet differentiable from $L^\infty(\Gamma_r \cup \Gamma_0)$ to $L^\infty(\Gamma_r \cup \Gamma_0)$ (cf. [5]). Since all other parts of A_q are linear and continuous in their respective spaces, in particular $G : L^\infty(\Gamma_r) \to L^\infty(\Gamma_r)$, A_q is clearly Fréchet differentiable from $W^{1,q}(\Omega)$ to $W^{1,q'}(\Omega)^*$, and its derivative at $\bar{y}$ in an arbitrary direction $y \in W^{1,q}(\Omega)$ is given by

$$\begin{aligned}\langle A_q'(\bar{y})y, v\rangle = \int_\Omega \kappa \nabla y \cdot \nabla v \, dx + 4\int_{\Gamma_r} (G\sigma|\bar{y}|^3 y) v \, ds \\ + 4\int_{\Gamma_0} \varepsilon\sigma|\bar{y}|^3 y v \, ds, \quad v \in W^{1,q'}(\Omega).\end{aligned} \tag{3.1}$$

By the same arguments, A_q is also twice continuously Fréchet differentiable, and the second derivative at $\bar{y}$ in arbitrary directions $y_1, y_2 \in W^{1,q}(\Omega)$ is given by

$$\langle A_q''(\bar{y})[y_1, y_2], v\rangle = 12\int_{\Gamma_r} (G\sigma|\bar{y}|\bar{y}\, y_1 y_2) v \, ds + 12\int_{\Gamma_0} \varepsilon\sigma|\bar{y}|\bar{y}\, y_1 y_2 \, v \, ds, \quad v \in W^{1,q'}(\Omega). \tag{3.2}$$

Notice that $A_q''(\bar{y})$ is clearly continuous from $W^{1,q}(\Omega) \times W^{1,q}(\Omega)$ to $W^{1,q'}(\Omega)^*$. Now, consider the operator equation

$$A_q'(\bar{y})y = w \quad \text{in} \quad W^{1,q'}(\Omega)^* \tag{3.3}$$

with a given $w \in W^{1,q'}(\Omega)^*$. Our goal is to show the existence of a unique solution to (3.3). In [17], an analogous equation in $H^1(\Omega)^*$ is investigated and is illustrated by means of a numerical example in which the Lax–Milgram lemma cannot be applied to derive existence of solutions because of the nonpositivity of G (cf. [17, section 4]). Instead of that, the Fredholm alternative is employed to prove existence and uniqueness (see also [18] for details). Here, we argue similarly, which is demonstrated in the following. First, we introduce a linear operator $F(\bar{y}) : L^\infty(\Gamma_r) \to W^{1,q'}(\Omega)^*$, defined by

$$\langle F(\bar{y})y, v\rangle := 4\int_{\Gamma_r} (G\sigma|\bar{y}|^3 y) v \, ds, \quad v \in W^{1,q'}(\Omega). \tag{3.4}$$

As already stated in section 2, the trace operator is continuous from $W^{1,q'}(\Omega)$ to $L^{r'}(\Gamma_r)$, $r' > 1$. Hence, thanks to $\bar{y} \in L^\infty(\Gamma_r)$, $F(\bar{y})$ is linear and continuous. Then, together with the Definition of B in Lemma 2.1, (3.3) is equivalent to

$$\big(B(\bar{\alpha}_0) + F(\bar{y})\tau_r\big)y = w, \tag{3.5}$$

where $\bar{\alpha}_0$ is defined by $\bar{\alpha}_0 := 4\varepsilon\sigma|\bar{y}|^3$ such that $\bar{\alpha}_0 \in L^\infty(\Gamma_0)$. Moreover, here and in the following, τ_r is considered as an operator from $W^{1,q}(\Omega)$ to $L^\infty(\Gamma_r)$. Now, since $\bar{y}(x) \geq \theta > 0$ a.e. on Γ_r, Lemma 2.1 is applicable such that

$$y = B(\bar{\alpha}_0)^{-1}(w - F(\bar{y})\tau_r y) = B(\bar{\alpha}_0)^{-1}w - B(\bar{\alpha}_0)^{-1}F(\bar{y})\tau_r y. \tag{3.6}$$

Applying τ_r to (3.6), we infer further

$$\big(I + \tau_r B(\bar{\alpha}_0)^{-1}F(\bar{y})\big)\tau_r y = \tau_r B(\bar{\alpha}_0)^{-1}w. \tag{3.7}$$

Let us now define a linear and continuous operator $\mathcal{F}(\bar{y}) : L^\infty(\Gamma_r) \to L^\infty(\Gamma_r)$ by

$$\mathcal{F}(\bar{y}) := \tau_r B(\bar{\alpha}_0)^{-1} F(\bar{y}), \tag{3.8}$$

and hence (3.7) is equivalent to

$$(I + \mathcal{F}(\bar{y}))\tau_r y = \tau_r B(\bar{\alpha}_0)^{-1} w \quad \text{in} \quad L^\infty(\Gamma_r). \tag{3.9}$$

We point out that, due to $q > N$, the trace operator τ_r is compact from $W^{1,q}(\Omega)$ to $L^\infty(\Gamma_r)$ (see [1]). Hence, $\mathcal{F}(\bar{y})$ is compact as well.

Assumption 3.1. *The operator $\mathcal{F}(\bar{y}) : L^\infty(\Gamma_r) \to L^\infty(\Gamma_r)$ does not admit the eigenvalue $\lambda = -1$.*

By virtue of Fredholm's theorem, Assumption 3.1 on the eigenvalue of $\mathcal{F}(\bar{y})$ immediately implies the existence and uniqueness of a solution $y \in W^{1,q}(\Omega)$ of the linearized equation (3.3) (see Theorem 3.1 below). Such a restriction has been considered in [15] which turns out to be particularly essential in order to establish the first- and second-order optimality conditions for (P). Therefore, we shall address a sufficient condition guaranteeing the case where -1 is not an eigenvalue of $\mathcal{F}(\bar{y})$. Recall that Friedrich-type and trace inequalities imply the existence of positive constants $C_F(\Omega), C_\tau(\Omega)$ independent of $y \in H^1(\Omega)$ such that

$$\begin{aligned} &\|y\|^2_{H^1(\Omega)} \le C_F(\Omega)(\|\nabla y\|^2_{L^2(\Omega)} + \|y\|^2_{L^2(\Gamma_0)}) && \forall y \in H^1(\Omega), \\ &\|y\|^2_{L^2(\Gamma_r)} \le C_\tau(\Omega)\|y\|^2_{H^1(\Omega)} && \forall y \in H^1(\Omega). \end{aligned} \tag{3.10}$$

In the upcoming proposition, we demonstrate that Assumption 3.1 is true in the case where the temperature difference on the interface surface Γ_r is small enough.

Proposition 3.1. *Let $\bar{u} \in L^2(\Omega_s)$ with $\bar{u}(x) \ge 0$ hold a.e. in Ω_s, and denote the associated state by $\bar{y} = \mathcal{G}(\bar{u})$. Notice that Lemma* 3.1 *implies $\bar{y} \ge \theta > 0$. Further, we define $\bar{y}_{\max,r} := \max_{x\in\Gamma_r}\{\bar{y}(x)\}$, $\bar{y}_{\min,r} := \min_{x\in\Gamma_r}\{\bar{y}(x)\}$ and $\bar{y}_{\min,0} := \min_{x\in\Gamma_0}\{\bar{y}(x)\}$. Under the assumption that*

$$\bar{y}^3_{\max,r} - \bar{y}^3_{\min,r} < \frac{1}{C_\tau(\Omega)\, C_F(\Omega)} \min\left\{\frac{1}{4\sigma}\kappa_{\min},\, \varepsilon_{\min}\bar{y}^3_{\min,0}\right\}, \tag{3.11}$$

the operator $\mathcal{F}(\bar{y}) : L^\infty(\Gamma_r) \to L^\infty(\Gamma_r)$ does not admit the eigenvalue $\lambda = -1$.

Proof. Let $z \in L^\infty(\Gamma_r)$ be a solution of $(I + \mathcal{F}(\bar{y}))z = 0$. Then, invoking the definition of $\mathcal{F}(\bar{y})$ (cf. (3.8)), the solution z satisfies

$$z = -\tau_r B(\alpha_0)^{-1} F(\bar{y}) z. \tag{3.12}$$

Now define $y_z \in W^{1,q}(\Omega)$ by $y_z := -B(\alpha_0)^{-1}F(\bar{y})z$ such that (3.12) implies $z = \tau_r y_z$. Hence, by construction, y_z solves

$$y_z = -B(\alpha_0)^{-1} F(\bar{y})\tau_r y_z. \tag{3.13}$$

Now, we demonstrate that $y_z = 0$, which immediately verifies the assertion. Equation (3.13) can also be written as $(B(\alpha_0) + F(\bar{y})\tau_r)y_z = 0$ which is, as noticed earlier in (3.5), equivalent to

$$A'_q(\bar{y})y_z = 0 \quad \text{in} \quad W^{1,q'}(\Omega)^*.$$

In view of (3.1), we may rewrite the above operator equation as

$$\int_\Omega \kappa\nabla y_z \cdot \nabla v \; dx + 4\int_{\Gamma_r} (G\sigma|\bar{y}|^3 y_z)v \; ds + 4\int_{\Gamma_0} \varepsilon\sigma|\bar{y}|^3 y_z v \; ds = 0 \quad \forall v \in W^{1,q'}(\Omega). \tag{3.14}$$

Setting $v = y_z$ in (3.14) leads to

$$\int_\Omega \kappa \nabla y_z \cdot \nabla y_z \, dx + 4 \int_{\Gamma_r} G(\sigma |\bar{y}|^3 y_z) y_z \, ds + 4 \int_{\Gamma_0} \varepsilon \sigma |\bar{y}|^3 y_z^2 \, ds = 0. \tag{3.15}$$

According to [21, Lemma 5], the operator G can be written as $G = I - H$, where $H \in \mathcal{B}(L^p(\Gamma_r))$ satisfies $\|H\|_{L^p(\Gamma_r)} \leq 1$ for all $1 \leq p \leq \infty$. By this decomposition and Lemma 3.1, the integral over Γ_r in (3.15) can be estimated as follows:

$$\begin{aligned} 4 \int_{\Gamma_r} G(\sigma |\bar{y}|^3 y_z) y_z \, ds &= 4 \int_{\Gamma_r} \big(\sigma |\bar{y}|^3 y_z^2 - H(\sigma |\bar{y}|^3 y_z) y_z\big) ds \\ &\geq 4\sigma (\bar{y}_{\min,r}^3 - \bar{y}_{\max,r}^3) \|y_z\|_{L^2(\Gamma_r)}^2. \end{aligned}$$

Due to $(\mathcal{A}_3)$ and $(\mathcal{A}_4)$, the above inequality and (3.15) imply

$$\begin{aligned} 0 &\geq \kappa_{\min} \|\nabla y_z\|_{L^2(\Omega)}^2 + 4\varepsilon_{\min} \sigma \bar{y}_{\min,0}^3 \|y\|_{L^2(\Gamma_0)}^2 + 4\sigma (\bar{y}_{\min,r}^3 - \bar{y}_{\max,r}^3) \|y_z\|_{L^2(\Gamma_r)}^2 \\ &\geq \frac{\min\{\kappa_{\min},\, 4\varepsilon_{\min} \sigma \bar{y}_{\min,0}^3\}}{C_F(\Omega)} \|y\|_{H^1(\Omega)}^2 + C_\tau(\Omega)\, 4\sigma (\bar{y}_{\min,r}^3 - \bar{y}_{\max,r}^3) \|y_z\|_{H^1(\Omega)}^2 \\ &= \left(\frac{\min\{\kappa_{\min},\, 4\varepsilon_{\min} \sigma \bar{y}_{\min,0}^3\}}{C_F(\Omega)} - C_\tau(\Omega)\, 4\sigma (\bar{y}_{\max,r}^3 - \bar{y}_{\min,r}^3) \right) \|y_z\|_{H^1(\Omega)}^2, \end{aligned} \tag{3.16}$$

where (3.10) was taken into account. Therefore, under the assumption (3.11), we come to the conclusion that $\|y_z\|_{H^1(\Omega)} = 0$, which completes the proof. □

Remark 3.1. It seems to be difficult to verify condition (3.11) in practice since it contains the unknown solution $\bar{y}$. However, if $\bar{y}$ is a locally optimal solution of (P), then (3.11) might hold since, in the context of crystal growth, small temperature gradients in the gas phase are desirable to improve the quality of the grown crystals (see [16] and the references therein). Thus the temperature differences on Γ_r are comparatively small in practice. Apart from this point of view, additional regularity assumptions such as Assumption 3.1 are typical for the Fréchet differentiability of solution operators associated with semilinear PDEs. Moreover, $\mathcal{F}(\bar{y})$ is a compact operator which is well known to possess only countably many eigenvalues so that Assumption 3.11 might be fulfilled in many cases.

With Assumption 3.1 at hand, we address the existence and uniqueness of the solution to (3.3) and the differentiability of the associated control-to-state operator. As pointed out previously, the assertion is an immediate result of Fredholm's theorem.

THEOREM 3.1. *Let $\bar{u} \in L^2(\Omega_s)$ with $\bar{u}(x) \geq 0$ a.e. in Ω_s, and denote the associated state by $\bar{y} = \mathcal{G}(\bar{u})$. Moreover, suppose that Assumption* 3.1 *holds true. Then, for every $w \in W^{1,q'}(\Omega)^*$, there exists a unique solution $y \in W^{1,q}(\Omega)$ to* (3.3) *that satisfies the estimate*

$$\|y\|_{W^{1,q}(\Omega)} \leq c \, \|w\|_{W^{1,q'}(\Omega)^*} \tag{3.17}$$

with a constant $c > 0$ independent of w. Hence, $A_q'(\bar{y})^{-1} \in \mathcal{B}(W^{1,q'}(\Omega)^*, W^{1,q}(\Omega))$ holds true.

Proof. Thanks to the compactness of $\mathcal{F}(\bar{y})$, the theory of Fredholm operators implies that either $\lambda = -1$ is one of countable many eigenvalues of $\mathcal{F}(\bar{y})$, or $I + \mathcal{F}(\bar{y})$ is continuously invertible. Hence, Assumption 3.1 ensures that $(I + \mathcal{F}(\bar{y}))^{-1} \in \mathcal{B}(L^\infty(\Gamma_r))$ such that

$$\tau_r y = (I + \mathcal{F}(\bar{y}))^{-1} \tau_r B(\bar{\alpha}_0)^{-1} w.$$

Inserting this in (3.6), we have

$$y = B(\bar{\alpha}_0)^{-1}(I - F(\bar{y})(I + \mathcal{F}(\bar{y}))^{-1}\tau_r B(\bar{\alpha}_0)^{-1})w. \tag{3.18}$$

Since Assumption 3.1 ensures that

$$\|(I + \mathcal{F}(\bar{y}))^{-1}\|_{\mathcal{B}(L^\infty(\Gamma_r))} < \infty,$$

(3.18) immediately implies (3.17). □

THEOREM 3.2. *Let $\bar{u} \in L^2(\Omega_s)$ with $\bar{u}(x) \geq 0$ a.e. in Ω_s. Furthermore, suppose that Assumption* 3.1 *is fulfilled. Then, there exists an open neighborhood $U(\bar{u})$ of $\bar{u}$ in $L^2(\Omega_s)$ such that $\mathcal{G} : L^2(\Omega_s) \to W^{1,q}(\Omega)$ is on $U(\bar{u})$ twice continuously Fréchet differentiable. Moreover, the first derivative of $\mathcal{G}$ at $\bar{u}$ in an arbitrary direction $u \in L^2(\Omega_s)$ is given by*

$$\mathcal{G}'(\bar{u})u = A_q'(\bar{y})^{-1}E_{q,s}\, u \tag{3.19}$$

with $\bar{y} = \mathcal{G}(\bar{u})$. The second derivative of $\mathcal{G}$ at $\bar{u}$ in arbitrary directions $u_1, u_2 \in L^2(\Omega_s)$ is given by

$$\mathcal{G}''(\bar{u})[u_1, u_2] = A_q'(\bar{y})^{-1}(-A_q''(\tilde{y})[y_1, y_2]), \tag{3.20}$$

where $A_q''(\bar{y})$ is defined as in (3.2) *and $y_i = \mathcal{G}'(\bar{u})u_i$, $i = 1, 2$.*

Proof. The proof follows standard arguments. First of all, let us introduce the operator $T : W^{1,q}(\Omega) \times L^2(\Omega_s) \to W^{1,q'}(\Omega)^*$, given by

$$T(y, u) := A_q(y) - E_{q,s}\, u - E_{q,0}\, \varepsilon\sigma y_0^4. \tag{3.21}$$

Further, we set $\bar{y} = \mathcal{G}(u)$ and hence, by the definition of the solution operator $\mathcal{G}$, $\bar{y} \in W^{1,q}(\Omega)$ is the unique solution of

$$A_q(\bar{y}) = E_{q,s}\, \bar{u} + E_{q,0}\, \varepsilon\sigma y_0^4 \text{ in } W^{1,q'}(\Omega)^*.$$

Thus, it holds that $T(\bar{y}, \bar{u}) = 0$. Moreover, since $A_q : W^{1,q}(\Omega) \to W^{1,q'}(\Omega)$ is twice continuously Fréchet differentiable, $T : W^{1,q}(\Omega) \times L^2(\Omega_s) \to W^{1,q'}(\Omega)^*$ is twice continuously Fréchet differentiable. By (3.21), $\partial_y T(\bar{y}, \bar{u}) : W^{1,q}(\Omega) \to W^{1,q'}(\Omega)^*$ is given by

$$\partial_y T(\bar{y}, \bar{u}) = A'(\bar{y}).$$

Therefore, Theorem 3.1 implies that $\partial_y T(\bar{y}, \bar{u})^{-1} \in \mathcal{B}(W^{1,q'}(\Omega)^*, W^{1,q}(\Omega))$. Thus, taking account of the implicit function theorem, we see there exists an open neighborhood $U(\bar{u})$ of $\bar{u}$ in $L^2(\Omega_s)$ such that the control-to-state operator $\mathcal{G} : L^2(\Omega_s) \to W^{1,q}(\Omega)$ is on $U(\bar{u})$ twice continuously Fréchet differentiable. The first derivative of $\mathcal{G}$ at $\bar{u}$ in an arbitrarily direction $u \in L^2(\Omega_s)$ is given by

$$\mathcal{G}'(\bar{u})u = -\partial_y T(\bar{y}, \bar{u})^{-1}\partial_u T(\bar{y}, \bar{u})u = A_q'(\bar{y})^{-1}E_{q,s}\, u. \tag{3.22}$$

Moreover, the second derivative of $\mathcal{G}$ at $\bar{u}$ in arbitrary directions $u_1, u_2 \in L^2(\Omega_s)$ is given by

$$\mathcal{G}''(\bar{u})[u_1, u_2] = -\partial_y T(\bar{y}, \bar{u})^{-1}\partial_{yy}^2 T(\bar{y}, \bar{u})[y_1, y_2] = A_q'(\bar{y})^{-1}(-A_q''(\bar{y})[y_1, y_2]),$$

where $y_i = \mathcal{G}'(\bar{u})u_i$, $i = 1, 2$. Notice that $\partial^2_{yu}T = \partial^2_{uu}T = 0$ and $\partial^2_{uy}T = 0$ were used for the computation of $\mathcal{G}''$. □

Remark 3.2. Notice that the additional assumption $\bar{u}(x) \geq 0$ a.e. in Ω_s is automatically fulfilled for all $u \in \mathcal{U}$, since $u_a(x) \geq 0$ a.e. in Ω_s.

In view of the definition of $A'_q(\bar{y})$ in (3.1) and formal integration by parts, $y := \mathcal{G}'(\bar{u})u$ and $\eta := \mathcal{G}''(\bar{u})[u_1, u_2]$, as given in (3.19) and (3.20), can be seen as solutions of linear PDEs. First, the equation $A'_q(\bar{y})y = E_{q,s}u$, which corresponds to (3.19), can be considered as the variational formulation of the following linear PDE:

$$\left\{\begin{aligned} -\operatorname{div}(\kappa_s \nabla y) &= u && \text{in } \Omega_s,\\ -\operatorname{div}(\kappa_g \nabla y) &= 0 && \text{in } \Omega_g,\\ \kappa_g\,(\partial_{n_r} y)_g - \kappa_s\,(\partial_{n_r} y)_s - 4\,G\sigma|\bar{y}|^3 y &= 0 && \text{on } \Gamma_r,\\ \kappa_s \partial_{n_0} y + 4\,\varepsilon\sigma|\bar{y}|^3 y &= 0 && \text{on } \Gamma_0. \end{aligned}\right. \tag{3.23}$$

Similarly, $A'_q(\bar{y})\eta = -A''_q(\bar{y})[y_1, y_2]$ is interpreted as a variational formulation of

$$\left\{\begin{aligned} -\operatorname{div}(\kappa_s \nabla y) &= 0 && \text{in } \Omega_s,\\ -\operatorname{div}(\kappa_g \nabla y) &= 0 && \text{in } \Omega_g,\\ \kappa_g\,(\partial_{n_r} y)_g - \kappa_s\,(\partial_{n_r} y)_s - 4\,G\sigma|\bar{y}|^3 y &= -12\,G\sigma|\bar{y}|\bar{y}\,y_1 y_2 && \text{on } \Gamma_r,\\ \kappa_s \partial_{n_0} y + 4\,\varepsilon\sigma|\bar{y}|^3 y &= -12\,\varepsilon\sigma|\bar{y}|\bar{y}\,y_1 y_2 && \text{on } \Gamma_0. \end{aligned}\right. \tag{3.24}$$

DEFINITION 3.1. *Let $\bar{u} \in L^2(\Omega_s)$, $\bar{u}(x) \geq 0$ a.e. in Ω_s, and $\bar{y} = \mathcal{G}(\bar{u})$ be given. Then, a function $y \in W^{1,q}(\Omega)$ is said to be a (weak) solution of* (3.23) *for $u \in L^2(\Omega_s)$ if it satisfies the following operator equation:*

$$A'_q(\bar{y})y = E_{q,s}\,u \quad \textit{in} \quad W^{1,q'}(\Omega)^*,$$

where $A'_q(\bar{y})$ is as defined in (3.1). *Moreover, $\eta \in W^{1,q}(\Omega)$ is the (weak) solution of* (3.24) *for given $y_1, y_2 \in W^{1,q}(\Omega)$ if it fulfills*

$$A'_q(\bar{y})\eta = -A''_q(\bar{y})[y_1, y_2] \quad \textit{in} \quad W^{1,q'}(\Omega)^*.$$

COROLLARY 3.1. *Suppose that $\bar{u} \in L^2(\Omega_s)$ is given with $\bar{u}(x) \geq 0$ a.e. in Ω_s, and assume that Assumption* 3.1 *is fulfilled. Then the first and second derivatives of $\mathcal{G}$ at $\bar{u}$ in direction $u \in L^2(\Omega_s)$ and directions $u_1, u_2 \in L^2(\Omega_s)$, respectively, are given by the solutions of* (3.23) *and* (3.24) *in the sense of Definition* 3.1.

4. Adjoint equation involving measures. In this section, we discuss the adjoint equation to (3.3), given by

$$A'_q(\bar{y})^* p = g \quad \text{in} \quad W^{1,q}(\Omega)^*, \tag{4.1}$$

where $A'_q(\bar{y})^* : W^{1,q'}(\Omega) \to W^{1,q}(\Omega)^*$ denotes the adjoint of $A'_q(\bar{y})$ and g is a given element of $W^{1,q}(\Omega)^*$. We already know from Theorem 3.1 that, under Assumption 3.1, $A'_q(\bar{y})$ is an isomorphism from $W^{1,q}(\Omega)$ to $W^{1,q'}(\Omega)^*$. Thus, the adjoint operator $A'_q(\bar{y})^* : W^{1,q'}(\Omega) \to W^{1,q}(\Omega)^*$ is in turn continuously invertible, and consequently, (4.1) admits a unique solution $p \in W^{1,q'}(\Omega)$, $q' = \frac{q}{q-1} < \frac{N}{N-1}$, due to $q > N$.

LEMMA 4.1. *Let $\bar{u} \in L^2(\Omega_s)$ with associated state $\bar{y} = \mathcal{G}(\bar{u})$ satisfy $\bar{u}(x) \geq 0$ a.e. in Ω_s. Furthermore, suppose that Assumption* 3.1 *is satisfied. Then, $A'_q(\bar{y})^{-*} \in \mathcal{B}(W^{1,q}(\Omega)^*, W^{1,q'}(\Omega))$ holds true.*

The concrete form of $A'_q(\bar{y})^*$ follows from

$$\begin{aligned}\langle A'_q(\bar{y})^* p, v\rangle_{(W^{1,q})^*,W^{1,q}} &= \langle p, A'_q(\bar{y})v\rangle_{W^{1,q'},(W^{1,q'})^*} = \langle A'_q(\bar{y})v, p\rangle_{(W^{1,q'})^*,W^{1,q'}}\\ &= \int_\Omega \kappa\nabla p\cdot\nabla v\,dx + 4\int_{\Gamma_r}(G\sigma|\bar{y}|^3 v)p\,ds + 4\int_{\Gamma_0}\varepsilon\sigma|\bar{y}|^3 pv\,ds, \quad v\in W^{1,q}(\Omega).\end{aligned}$$

As demonstrated in the following, we are allowed to insert regular Borel measures as inhomogeneity in (4.1). To this end, define by $\mathcal{M}(\overline{\Omega})$ the space of all regular Borel measures on the compact set $\overline{\Omega}$. By the Riesz–Radon theorem (cf. [4]), it is well known that the dual space $\mathcal{C}(\overline{\Omega})^*$ can be isometrically identified with $\mathcal{M}(\overline{\Omega})$ with respect to the duality pairing

$$\langle \mu, \varphi\rangle_{\mathcal{C}(\overline{\Omega})^*,\mathcal{C}(\overline{\Omega})} := \int_{\overline{\Omega}}\varphi\,d\mu, \quad \phi\in\mathcal{C}(\overline{\Omega}),\ \mu\in\mathcal{M}(\overline{\Omega}).$$

According to this, we associate with every $\mu\in\mathcal{M}(\overline{\Omega})$ an element of $W^{1,q}(\Omega)^*$, denoted by $\tilde{\mu}$, by setting

$$\langle\tilde{\mu}\,,\,v\rangle_{W^{1,q}(\Omega)^*,W^{1,q}(\Omega)} = \int_{\overline{\Omega}} v\,d\mu, \quad v\in W^{1,q}(\Omega). \tag{4.2}$$

Notice that the right-hand side in (4.2) clearly defines an element of $W^{1,q}(\Omega)^*$ since $W^{1,q}(\Omega)\hookrightarrow\mathcal{C}(\overline{\Omega})$. Hence, for a given $\mu\in\mathcal{M}(\overline{\Omega})$, the operator equation

$$A'_q(\bar{y})^* p = \tilde{\mu} \quad \text{in} \quad W^{1,q}(\Omega)^* \tag{4.3}$$

is equivalent to

$$\int_\Omega \kappa\nabla p\cdot\nabla v\,dx + 4\int_{\Gamma_r}(G\sigma|\bar{y}|^3 v)p\,ds + 4\int_{\Gamma_0}\varepsilon\sigma|\bar{y}|^3 pv\,ds = \int_{\overline{\Omega}} v\,d\mu \quad \forall\, v\in W^{1,q}(\Omega). \tag{4.4}$$

As mentioned in section 2, the trace operator is continuous from $W^{1,q'}(\Omega)$ to $L^{r'}(\Gamma_r)$, $r' = \frac{(N-1)q}{(N-1)q-N} > 1$. Moreover, $v\in W^{1,q}(\Omega)$ clearly implies $v\in L^r(\Gamma_r)$ due to the continuous embedding $W^{1,q}(\Omega)\hookrightarrow\mathcal{C}(\overline{\Omega})$. Hence, if we consider G as an operator from $L^r(\Gamma_r)$ to $L^r(\Gamma_r)$, we obtain

$$\int_{\Gamma_r}(G\sigma|\bar{y}|^3 v)p\,ds = \int_{\Gamma_r}\sigma|\bar{y}|^3(G^* p)v\,ds,$$

where $G^*: L^{r'}(\Gamma_r)\to L^{r'}(\Gamma_r)$ is the adjoint of G, i.e., $G^* = \varepsilon(I-(1-\varepsilon)K^*)^{-1}(I-K^*)$ (cf. Definition 1.1). Notice in this context that K is formally self-adjoint due to the symmetry of its kernel. In view of this and formal integration by parts, (4.3) and (4.4), respectively, can be considered as a variational formulation of the following linear PDE with measure data on the right-hand side:

$$\begin{cases} -\mathrm{div}(\kappa_s\,\nabla p) = \mu_{|\Omega_s} & \text{in } \Omega_s,\\ -\mathrm{div}(\kappa_g\,\nabla p) = \mu_{|\Omega_g} & \text{in } \Omega_g,\\ \kappa_g\left(\dfrac{\partial p}{\partial_{n_r}}\right)_g - \kappa_s\left(\dfrac{\partial p}{\partial_{n_r}}\right)_s - 4\,\sigma|\bar{y}|^3 G^* p = \mu_{|\Gamma_r} & \text{on } \Gamma_r,\\ \kappa_s\partial_{n_0}p + 4\,\varepsilon\sigma|\bar{y}|^3 p = \mu_{|\Gamma_0} & \text{on } \Gamma_0,\end{cases} \tag{4.5}$$

where $\mu_{|\Omega_g}$, $\mu_{|\Omega_s}$, $\mu_{|\Gamma_r}$, and $\mu_{|\Gamma_0}$ denote the restrictions of μ on Ω_g, Ω_s, Γ_r, and Γ_0, respectively. In other words, $\mu \in \mathcal{M}(\overline{\Omega})$ is decomposed into $\mu = \mu_{|\Omega_g} + \mu_{|\Omega_s} + \mu_{|\Gamma_r} + \mu_{|\Gamma_0}$, where $\mu_{|\Omega_g}$, $\mu_{|\Omega_s}$, $\mu_{|\Gamma_r}$, and $\mu_{|\Gamma_0}$ are Borel measures concentrated on Ω_g, Ω_s, Γ_r, and Γ_0.

Definition 4.1. *Let $\bar{y} \in W^{1,q}(\Omega)$ be given. Then, a function $p \in W^{1,q'}(\Omega)$, $q' < \frac{N}{N-1}$, is said to be a (weak) solution of* (4.5) *if it satisfies the operator equation* (4.3)*, which is equivalent to the weak formulation* (4.4)*.*

Clearly, Lemma 4.1 implies that there is a solution of (4.5) in the sense of Definition 4.1. Furthermore, since the embedding $W^{1,q}(\Omega) \hookrightarrow \mathcal{C}(\overline{\Omega})$ is continuous and dense, $\mathcal{M}(\overline{\Omega})$ is continuously embedded in $W^{1,q}(\Omega)^*$, and hence

$$\|\tilde{\mu}\|_{W^{1,q}(\Omega)^*} \leq c\,\|\mu\|_{\mathcal{M}(\overline{\Omega})}.$$

Consequently one obtains the following result.

Theorem 4.1. *Let $\bar{u} \in L^2(\Omega_s)$ with $\bar{u}(x) \geq 0$ a.e. in Ω_s, and let the associated state be denoted by $\bar{y} = \mathcal{G}(\bar{u}) \in W^{1,q}(\Omega)$. Furthermore, suppose that Assumption* 3.1 *is satisfied. Then, $A_q(\bar{y})^{-*} \in \mathcal{B}(W^{1,q}(\Omega)^*, W^{1,q'}(\Omega))$, and consequently, for every $\mu \in \mathcal{M}(\overline{\Omega})$, there exists a unique solution $p \in W^{1,q'}(\Omega)$ of* (4.5) *in the sense of Definition* 4.1 *that satisfies*

$$\|p\|_{W^{1,q'}(\Omega)} \leq c\,\|\mu\|_{\mathcal{M}(\overline{\Omega})}$$

with a constant $c > 0$ independent of μ.

5. First-order necessary optimality conditions for (P).

Before establishing KKT-type optimal conditions for (P), we briefly address the existence of an optimal solution. Clearly, under the assumption that there exists a feasible control $\bar{u}$ of (P), standard arguments imply the existence of at least one (global) optimum (cf. also [15, Theorem 5.2]). Due to the nonlinearities in the state equation, uniqueness of the optimal solution can certainly not be expected. Let us now introduce the notion of local optima.

Definition 5.1. *A feasible control $\bar{u}$ of* (P) *is called a local solution for* (P) *if there exists a positive real number ε such that $f(\bar{u}) \leq f(u)$ holds for all feasible $u \in L^2(\Omega_s)$ with $\|u - \bar{u}\|_{L^2(\Omega_s)} \leq \varepsilon$.*

Throughout this section, let $\bar{u} \in \mathcal{U}$ be a local solution of (P), and assume that *Assumption* 3.1 *is fulfilled at $\bar{u}$.* Notice that everything that follows also holds for a global optimum of (P). To apply the KKT theory, the existence of an interior (Slater) point with respect to the state constraints in (1.1) has to be assumed. This assumption is referred to as the "linearized Slater condition."

Definition 5.2. *Let $\bar{u} \in \mathcal{U}$ be a local solution of* (P)*, and assume that Assumption* 3.1 *is fulfilled at $\bar{u}$. We say that $\bar{u} \in \mathcal{U}$ satisfies the linearized Slater condition for* (P) *if there exists an interior point $u_0 \in \mathcal{U}$ such that*

$$\begin{aligned} y_a(x) + \delta \leq \mathcal{G}(\bar{u})(x) + \mathcal{G}'(\bar{u})(u_0 - \bar{u})(x) &\leq y_b(x) - \delta \qquad && \forall\, x \in \overline{\Omega}_g, \\ \mathcal{G}(\bar{u})(x) + \mathcal{G}'(\bar{u})(u_0 - \bar{u})(x) &\leq y_{\max}(x) - \delta \qquad && \forall\, x \in \overline{\Omega}_s, \end{aligned}$$

with a fixed $\delta > 0$.

Notice that $\bar{u} \in \mathcal{U}$ automatically satisfies $\bar{u}(x) \geq 0$ a.e. in Ω_s (cf. Remark 3.2) such that Assumption 3.1 implies that $\mathcal{G}'(\bar{u}) : L^2(\Omega_s) \to W^{1,q}(\Omega)$ is well defined.

Definition 5.3. *The Lagrange functional* $\mathcal{L} : \mathcal{U} \times \mathcal{M}(\overline{\Omega}_s) \times \mathcal{M}(\overline{\Omega}_g) \times \mathcal{M}(\overline{\Omega}_g) \to \mathbb{R}$ *for* (P) *is given by*

$$\mathcal{L}(u,\mu) = f(u) + \int_{\overline{\Omega}_s} (\mathcal{G}(u) - y_{\max}) d\mu_s + \int_{\overline{\Omega}_g} (y_a - \mathcal{G}(u)) d\mu_g^a + \int_{\overline{\Omega}_g} (\mathcal{G}(u) - y_b) d\mu_g^b,$$

with $\mu = (\mu_s, \mu_g^a, \mu_g^b)$.

Since $\mathcal{G}$ is twice continuously Fréchet differentiable at $\bar{u}$ (see Theorem 3.2), it is straightforward to see that f is twice continuously Fréchet differentiable at $\bar{u}$, and its derivative at $\bar{u} \in L^2(\Omega_s)$ in an arbitrary direction $u \in L^2(\Omega_s)$ is given by

$$f'(\bar{u})u = \int_{\Omega_g} (\nabla \mathcal{G}(\bar{u}) - z) \cdot \nabla \mathcal{G}'(\bar{u}) u \, dx + \beta \int_{\Omega_s} \bar{u}\, u \, dx.$$

Due to $\mathcal{G}(\bar{u}) \in W^{1,q}(\Omega)$ and $z \in L^2(\Omega_g)^N$, the first addend defines an element of $W^{1,q}(\Omega)^*$ such that the linear and continuous operator $L : W^{1,q}(\Omega) \to W^{1,q}(\Omega)^*$ exists with

$$\langle L\bar{y}, v\rangle := \int_{\Omega_g} (\nabla \bar{y} - z) \cdot \nabla v \, dx, \quad v \in W^{1,q}(\Omega),$$

where $\bar{y} = \mathcal{G}(\bar{u}) \in W^{1,q}(\Omega)$. With this setting, $f'(\bar{u})u = \langle L\bar{y}, \mathcal{G}'(\bar{u})u\rangle + \beta(\bar{u}, u)$. Notice that since f and $\mathcal{G}$ are continuously Fréchet differentiable at $\bar{u}$, $\mathcal{L}$ is continuously Fréchet differentiable at $\bar{u}$ such that the following definition makes sense.

Definition 5.4. *Let* $\bar{u} \in \mathcal{U}$ *be a local solution of* (P), *and suppose that Assumption* 3.1 *is fulfilled. Then,* $\mu_s \in \mathcal{M}(\overline{\Omega}_s)$, $\mu_g^a \in \mathcal{M}(\overline{\Omega}_g)$, *and* $\mu_g^b \in \mathcal{M}(\overline{\Omega}_g)$ *are said to be Lagrange multipliers associated with the state constraints in* (P) *if it holds that*

$$\partial_u \mathcal{L}(\bar{u},\mu)(u - \bar{u}) \geq 0 \quad \forall\, u \in \mathcal{U}, \tag{5.1}$$

$$\mu_s \geq 0, \quad \mu_g^a \geq 0, \quad \mu_g^b \geq 0, \tag{5.2}$$

$$\int_{\overline{\Omega}_s} (\mathcal{G}(\bar{u}) - y_{\max}) d\mu_s = \int_{\overline{\Omega}_g} (y_a - \mathcal{G}(\bar{u})) d\mu_g^a = \int_{\overline{\Omega}_g} (\mathcal{G}(\bar{u}) - y_b) d\mu_g^b = 0, \tag{5.3}$$

where we again set $\mu = (\mu_s, \mu_g^a, \mu_g^b)$.

Notice that if $\nu \in \mathcal{M}(\overline{\Omega})$, then we write

$$\nu \geq 0 \quad \Leftrightarrow \quad \int_{\overline{\Omega}} y \, d\nu \geq 0 \quad \forall\, y \in \{y \in \mathcal{C}(\overline{\Omega}) \,|\, y(x) \geq 0 \,\forall\, x \in \overline{\Omega}\}.$$

The following theorem states the first-order necessary optimality conditions for (P), i.e., the existence of Lagrange multipliers in the sense of Definition 5.4. The proof can be found, for instance, in [10].

Theorem 5.1. *Let* $\bar{u}$ *be a locally optimal solution of* (P) *satisfying the linearized Slater condition. Furthermore, let Assumption* 3.1 *be satisfied. Then, there exist corresponding Lagrange multipliers* $(\mu_s, \mu_g^a, \mu_g^b) \in \mathcal{M}(\overline{\Omega}_s) \times \mathcal{M}(\overline{\Omega}_g) \times \mathcal{M}(\overline{\Omega}_g)$ *according to Definition* 5.4 *such that* (5.1)–(5.3) *are satisfied.*

Next, let us transform (5.1)–(5.3) into the optimality system of (P) by introducing the adjoint equation. First, by the definition of $\mathcal{L}$ and (3.22), (5.1) is equivalent to

$$\begin{aligned}\langle A_q'(\bar{u})^{-*}\big(L\bar{y} + \tilde{\mu}_s - \tilde{\mu}_g^a + \tilde{\mu}_g^b\big), E_{q,s}(u - \bar{u})\rangle_{W^{1,q'}(\Omega), W^{1,q'}(\Omega)^*}& \\ + (\beta\, \bar{u}\,,\, u - \bar{u})_{L^2(\Omega_s)} \geq 0 \quad \forall\, u \in \mathcal{U},&\end{aligned} \tag{5.4}$$

where $\tilde{\mu}_s$, $\tilde{\mu}_g^a$, and $\tilde{\mu}_g^b$ denote the elements of $W^{1,q}(\Omega)^*$ associated with μ_s, μ_g^a, and μ_g^b in the sense of (4.2). Consider now the operator equation

$$A_q'(\bar{y})^* p = L\bar{y} + \tilde{\mu}_s - \tilde{\mu}_g^a + \tilde{\mu}_g^b \quad \text{in} \quad W^{1,q}(\Omega)^* \tag{5.5}$$

which is equivalent to

(5.6)
$$\begin{aligned}\int_\Omega \kappa \nabla p \cdot \nabla v\,dx &+ 4\int_{\Gamma_r} (G\sigma|\bar{y}|^3 v)p\,ds + 4\int_{\Gamma_0} \varepsilon\sigma|\bar{y}|^3 pv\,ds \\ &= \int_{\Omega_g} (\nabla\bar{y} - z)\cdot\nabla v\,dx + \int_{\overline{\Omega}_s} v\,d\mu_s - \int_{\overline{\Omega}_g} v\,d\mu_g^a + \int_{\overline{\Omega}_g} v\,d\mu_g^b \quad \forall\, v \in W^{1,q}(\Omega)\end{aligned}$$

(cf. (4.3) and (4.4)). As in the case of (4.5), (5.6) can be considered as the variational formulation of

(5.7)
$$\begin{cases} -\operatorname{div}(\kappa_g \nabla p) = -\Delta\bar{y} + \operatorname{div} z + (\mu_g^b - \mu_g^a)_{|\Omega_g} & \text{in } \Omega_g, \\ -\operatorname{div}(\kappa_s \nabla p) = \mu_{s|\Omega_s} & \text{in } \Omega_s, \\ \kappa_g \left(\dfrac{\partial p}{\partial_{n_r}}\right)_g - \kappa_s \left(\dfrac{\partial p}{\partial_{n_r}}\right)_s - 4\sigma|\bar{y}|^3 G^\star p = -\dfrac{\partial \bar{y}}{\partial n_r} + z\cdot n_r \\ \qquad\qquad + (\mu_g^b - \mu_g^a + \mu_s)_{|\Gamma_r} & \text{on } \Gamma_r, \\ \kappa_s \dfrac{\partial p}{\partial n_0} + 4\varepsilon\sigma|\bar{y}|^3 p = \mu_{s|\Gamma_0} & \text{on } \Gamma_0. \end{cases}$$

Again, the multipliers are decomposed into their restrictions on Ω_s, Ω_g, Γ_r, and Γ_0, respectively. Analogously to Definition 4.1, we define solutions to (5.7).

Definition 5.5. *A function $p \in W^{1,q'}(\Omega)$ is said to be the weak solution of* (5.7) *if it satisfies* (5.5) *and* (5.6)*, respectively.*

Clearly, thanks to Lemma 4.1, there exists a unique solution of (5.7) in the sense of Definition 5.5 (cf. Theorem 4.1). Using the definition of p and (5.4), (5.1) can be transformed into

$$\frac{\partial\mathcal{L}}{\partial u}(\bar{u},\mu)(u-\bar{u}) = \int_{\Omega_s} (p + \beta\bar{u})(u-\bar{u})\,dx \geq 0 \quad \forall\, u \in \mathcal{U}. \tag{5.8}$$

By standard arguments, a pointwise evaluation of this equation implies

$$\bar{u} = \mathcal{P}_{ad}\Big\{-\frac{1}{\beta}p(x)\Big\}, \tag{5.9}$$

where $\mathcal{P}_{ad} : L^2(\Omega_s) \to L^2(\Omega_s)$ denotes the pointwise projection operator on the admissible set $\mathcal{U}$. In this way, we find the following theorem that states the first-order necessary optimality conditions for (P).

Theorem 5.2 (first-order necessary optimality conditions for (P)). *Let $\bar{u} \in L^2(\Omega_s)$ be an optimal solution of* (P) *with the associated state $\bar{y} = \mathcal{G}(\bar{u}) \in W^{1,q}(\Omega)$, $q > N$. Suppose further that $\bar{u}$ satisfies Assumption* 3.1 *and the linearized Slater conditions. Then, there exist an adjoint state $p \in W^{1,q'}(\Omega)$, $q' < \frac{N}{N-1}$, and Lagrange multipliers $\mu_s \in \mathcal{M}(\overline{\Omega}_s)$, $\mu_g^a \in \mathcal{M}(\overline{\Omega}_g)$, and $\mu_g^b \in \mathcal{M}(\overline{\Omega}_g)$ such that the following relations are satisfied:*

- *The state equation* (SL) *in the sense of Definition* 2.1,
- *the adjoint equation* (5.7) *in the sense of Definition* 5.5,

- *the projection formula* (5.9),
- *the nonnegativity of the Lagrange multipliers* (5.2), *and*
- *the complementary slackness conditions* (5.3).

It is straightforward to see that, if $u_a, u_b \in W^{1,q'}(\Omega_s)$, then $\mathcal{P}_{ad}$ is continuous from $W^{1,q'}(\Omega_s)$ to $W^{1,q'}(\Omega_s)$ such that the following regularity result for the optimal control is obtained.

Remark 5.1. If $u_a, u_b \in W^{1,q'}(\Omega_s)$, then the optimal control $\bar{u}$ is a function in $W^{1,q'}(\Omega_s)$, $q' < \frac{N}{N-1}$.

6. Second-order sufficient optimality conditions for (P).

In the following, we present second-order sufficient optimality conditions for (P) guaranteeing local optimality with respect to the $L^2(\Omega)$ topology. The investigation of second-order sufficient optimality conditions for semilinear control problems with pointwise state constraints was originally undertaken by Casas, Tröltzsch, and Unger in [8]. They suggested second-order optimality conditions that deal with strongly active sets. However, owing to the presence of the two-norm discrepancy, the result only provides sufficient optimality conditions for local solutions with respect to $L^\infty(\Omega)$ topology. Later on, Casas, De Los Reyes, and Tröltzsch [7] modified this result and arrived at sufficient conditions that are in some sense less restrictive than the original conditions. In particular, under certain assumptions, these conditions ensure the existence of local solutions in $L^2(\Omega)$. The result is, however, not directly applicable for (P) since here we deal with a nonmonotone operator G, and the objective functional in (P) is different from that considered in [7]. However, thanks to Theorem 3.2, we obtain analogous second-order sufficient conditions for the existence of local solutions to (P) in $L^2(\Omega_s)$. At this point, let us underline that the proof of Theorem 6.1 below is basically analogous to the second-order analysis of [7], which has only to be slightly modified due to the gradient within the objective functional of (P). In view of this, the corresponding analysis is therefore presented in a rather concise form.

DEFINITION 6.1. *Let $\bar{u} \in \mathcal{U}$ be a feasible control of* (P) *with the associated state $\mathcal{G}(\bar{u}) = \bar{y}$. We assume that there exist $\mu_g^a, \mu_g^b \in \mathcal{M}(\overline{\Omega}_g)$, $\mu_s \in \mathcal{M}(\overline{\Omega}_s)$, and $p \in W^{1,q'}(\Omega)$, $1 \le q' \le N/(N-1)$, satisfying* (5.1)–(5.3) *and* (5.7).

(i) *The convex, closed subset $\mathcal{H}_{\bar{u}} \subset L^2(\Omega_s)$ is given by*

$$\mathcal{H}_{\bar{u}} := \left\{ h \in L^2(\Omega_s) \mid h(x) = \begin{cases} \ge \; 0 & \textit{if} \quad \bar{u}(x) = u_a(x), \\ \le \; 0 & \textit{if} \quad \bar{u}(x) = u_b(x). \end{cases} \right\}$$

(ii) *The subset $\mathcal{C}_{\bar{u}} \subset \mathcal{H}_{\bar{u}}$ is defined as follows:*

$$\mathcal{C}_{\bar{u}} = \{ h \in \mathcal{H}_{\bar{u}} \mid h \textit{ satisfies } (6.1), (6.2), \textit{ and } (6.3)\},$$

$$h(x) = 0 \quad \textit{if} \quad p(x) + \beta \bar{u}(x) \ne 0, \tag{6.1}$$

$$y_h(x) = \begin{cases} \ge \; 0 & \textit{if} \quad \bar{y}(x) = y_a(x), \; x \in \overline{\Omega}_g, \\ \le \; 0 & \textit{if} \quad \bar{y}(x) = y_b(x), \; x \in \overline{\Omega}_g, \\ \le \; 0 & \textit{if} \quad \bar{y}(x) = y_{\max}(x), \; x \in \overline{\Omega}_s, \end{cases} \tag{6.2}$$

$$\int_{\bar{\Omega}_g} y_h \, d\mu_g^a = \int_{\bar{\Omega}_g} y_h \, d\mu_g^b = \int_{\bar{\Omega}_s} y_h \, d\mu_s = 0, \tag{6.3}$$

where $y_h = S'(\bar{u})h$.

(iii) *We say that $\bar{u}$ satisfies the second-order sufficient condition* (SSC) *if*

$$\text{(SSC)} \qquad \frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u}, \mu)h^2 > 0$$

holds true for every $h \in \mathcal{C}_{\bar{u}} \setminus \{0\}$.

Theorem 6.1 (second-order sufficient optimality conditions for (P)). *Let $\bar{u} \in \mathcal{U}$ be a feasible control of* (P) *and let Assumption* 3.1 *be fulfilled. Furthermore, assume that there exist $\mu_g^a, \mu_g^b \in \mathcal{M}(\overline{\Omega}_g)$, $\mu_s \in \mathcal{M}(\overline{\Omega}_s)$, and $p \in W^{1,q'}(\Omega)$, $1 \le q' \le N/(N-1)$, satisfying* (5.1)–(5.3) *and* (5.7). *If $\bar{u}$ additionally satisfies* (SSC), *then there exist positive real numbers ε and δ such that*

$$f(\bar{u}) + \frac{\delta}{2}\|u - \bar{u}\|^2_{L^2(\Omega_s)} \le f(u)$$

holds true for every feasible control u of (P) *with $\|u - \bar{u}\|_{L^2(\Omega_s)} < \varepsilon$. Hence, $\bar{u}$ is a local solution of* (P) *according to Definition* 5.1.

Proof. As already mentioned above, generally speaking, the proof completely follows the lines of [7]. With a modification of the arguments in [7] the gradient-type objective functional can also be incorporated into the analysis, as we will see in the following. Let us start by assuming the contrary: There exists a sequence $\{u_k\}_{k=1}^\infty \subset L^2(\Omega_s)$ of feasible controls of (P) such that

$$(6.4) \qquad f(\bar{u}) + \frac{1}{k}\|u_k - \bar{u}\|^2_{L^2(\Omega_s)} > f(u_k) \;\forall k \in \mathbb{N} \quad \text{and} \quad \lim_{k\to\infty} \|u_k - \bar{u}\|_{L^2(\Omega_s)} = 0.$$

We define $h_k := \frac{1}{a_k}(u_k - \bar{u})$ with $a_k := \|u_k - \bar{u}\|_{L^2(\Omega_s)}$. Thus, $\|h_k\|_{L^2(\Omega_s)} = 1$ holds for all $k \in \mathbb{N}$. For this reason, there exists a subsequence denoted w.l.o.g. again by $\{h_k\}_{k=1}^\infty$, which converges weakly in $L^2(\Omega_s)$ to some $\bar{h} \in L^2(\Omega_s)$, i.e., $h_k \rightharpoonup \bar{h}$ as $k \to \infty$. Using the same arguments as in [7, section 4], one proves that

$$(6.5) \qquad \frac{\partial \mathcal{L}}{\partial u}(\bar{u}, \mu)\bar{h} = 0.$$

The underlying argument is based on the mean value theorem, which is also applicable here since Theorem 3.2 guarantees that $\mathcal{G}$ is twice continuously differentiable in a neighborhood around $\bar{u}$. Moreover, it is completely analogous to [7, section 4] to show that $\bar{h} \in \mathcal{C}_{\bar{u}}$, i.e., $\bar{h}$ belongs to $\mathcal{H}_{\bar{u}}$ and satisfies conditions (6.1)–(6.3); cf. Definition 6.1. The corresponding argument relies on the compactness of $\mathcal{G}'(\bar{u})$ when considered with range in $\mathcal{C}(\overline{\Omega})$, which holds due to the continuous Fréchet differentiability of $\mathcal{G} : L^2(\Omega_s) \to W^{1,q}(\Omega)$ and the compact embedding $W^{1,q}(\Omega) \hookrightarrow \mathcal{C}(\overline{\Omega})$ for $q > N$.

From (6.5) and $\bar{h} \in \mathcal{C}_{\bar{u}}$, it follows that $\bar{h} = 0$. For the reader's convenience, this is demonstrated in more detail in the following since the arguments of [7] have to be slightly modified in the case of a gradient-type objective functional as in (P). Due to Theorem 3.2 and the convergence $u_k \to \bar{u}$ by construction, there is an open ball $B_r(\bar{u}) \subset L^2(\Omega_s)$, where $\mathcal{G}$ is twice continuously differentiable. Hence there exists a point $z_k \in L^2(\Omega_s)$ between u_k and $\bar{u}$ such that

$$\mathcal{L}(u_k, \mu) = \mathcal{L}(\bar{u}, \mu) + a_k \frac{\partial \mathcal{L}}{\partial u}(\bar{u}, \mu)h_k + \frac{a_k^2}{2}\frac{\partial^2 \mathcal{L}}{\partial u^2}(z_k, \mu)h_k^2 \quad \forall\, k \ge k_0$$

for sufficiently large $k_0 > 0$. By rearranging and dividing by $a_k^2/2$, the above equation is equivalent to

$$\begin{aligned}
2\frac{\partial \mathcal{L}}{\partial u}(\bar{u},\mu)(u_k-\bar{u}) + \frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u},\mu)h_k^2 =& \frac{2}{a_k^2}\{\mathcal{L}(u_k,\mu) - \mathcal{L}(\bar{u},\mu)\} \\
&+ \left[\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u},\mu) - \frac{\partial^2 \mathcal{L}}{\partial u^2}(z_k,\mu)\right]h_k^2 \quad \forall\, k \geq k_0.
\end{aligned} \tag{6.6}$$

Furthermore, the feasibility of u_k for (P) and the positivity of μ imply that $\mathcal{L}(u_k,\mu) \leq f(u_k)$, and hence from (6.4) we infer that

$$\mathcal{L}(u_k,\mu) \leq f(u_k) < f(\bar{u}) + \frac{1}{k}\|u_k-\bar{u}\|_{L^2(\Omega_s)}^2 = \mathcal{L}(\bar{u},\mu) + \frac{1}{k}\|u_k-\bar{u}\|_{L^2(\Omega_s)}^2 \quad \forall\, k \in \mathbb{N},$$

where we used the complementary slackness conditions (5.3). Hence, since $u_k \in \mathcal{U}$ and $\|h_k\|_{L^2(\Omega_s)} = 1$ for all $k \in \mathbb{N}$, the latter equality, together with (5.1) and (6.6), implies that

$$\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u},\mu)h_k^2 < \frac{1}{k} + \left\|\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u},\mu) - \frac{\partial^2 \mathcal{L}}{\partial u^2}(z_k,\mu)\right\|_{\mathcal{B}^2(L^2(\Omega_s))} \qquad \forall\, k \geq k_0, \tag{6.7}$$

where $\mathcal{B}^2(L^2(\Omega_s))$ denotes the space of bounded bilinear forms from $L^2(\Omega_s) \times L^2(\Omega_s)$ to $\mathbb{R}$. Notice that $\frac{\partial^2 \mathcal{L}}{\partial u^2}(\cdot,\mu)$ is continuous from $L^2(\Omega_s)$ to $\mathcal{B}^2(L^2(\Omega_s))$, and hence since $\lim_{k\to\infty} z_k = \bar{u}$ in $L^2(\Omega_s)$, the right-hand side of (6.7) converges to zero as $k \to \infty$. To show the convergence of the left-hand side in (6.7), we argue as follows: For each $k \in \mathbb{N}$, we set $y_k := \mathcal{G}'(\bar{u})h_k$ and $w_k := \mathcal{G}''(\bar{u})h_k^2$, and hence

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u},\mu)h_k^2 = \|\nabla y_k\|_{L^2(\Omega_g)}^2 + (\nabla\bar{y} - z, \nabla w_k)_{L^2(\Omega_g)} + \beta\|h_k\|_{L^2(\Omega_s)}^2 \\
+ \int_{\bar{\Omega}_s} w_k\, d\mu_s + \int_{\bar{\Omega}_g} w_k\, d\mu_g^b - \int_{\bar{\Omega}_g} w_k\, d\mu_g^a.
\end{aligned} \tag{6.8}$$

Obviously, since $\mathcal{G}'(\bar{u})$ is continuous and linear from $L^2(\Omega_s)$ to $W^{1,q}(\Omega)$ and $h_k \rightharpoonup \bar{h}$ in $L^2(\Omega_s)$, one finds

$$\nabla y_k \rightharpoonup \nabla y_{\bar{h}} \text{ in } L^2(\Omega) \text{ as } k \to \infty, \tag{6.9}$$

with $y_{\bar{h}} = \mathcal{G}'(\bar{u})\bar{h}$. Moreover, since $w_k = \mathcal{G}''(\bar{u})h_k^2 = -(A_q'(\bar{y}))^{-1}A_q''(\bar{y})[y_k,y_k]$ (see Theorem 3.2) and $y_k \to y_{\bar{h}}$ in $\mathcal{C}(\overline{\Omega})$ due to the compactness of $W^{1,q}(\Omega) \hookrightarrow \mathcal{C}(\overline{\Omega})$, we have

$$w_k \to w_{\bar{h}} := \mathcal{G}''(\bar{u})\bar{h}^2 \quad \text{in } W^{1,q}(\Omega).$$

Thus, using the weak lower semicontinuity of the norm, we arrive at

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u},\mu)\bar{h}^2 =& \|\nabla y_{\bar{h}}\|_{L^2(\Omega_g)}^2 + (\nabla\bar{y} - z, \nabla w_{\bar{h}})_{L^2(\Omega_g)} + \beta\|\bar{h}\|_{L^2(\Omega_s)}^2 \\
&+ \int_{\bar{\Omega}_s} w_{\bar{h}} d\mu_s - \int_{\bar{\Omega}_g} w_{\bar{h}} d\mu_g^a + \int_{\bar{\Omega}_g} w_{\bar{h}} d\mu_g^b \\
\leq& \liminf_{k\to\infty} \frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u},\mu)h_k^2 \\
\leq& \lim_{k\to\infty} \left\{\frac{1}{k} + \left\|\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u},\mu) - \frac{\partial^2 \mathcal{L}}{\partial u^2}(z_k,\mu)\right\|_{\mathcal{B}^2(L^2(\Omega_s))}\right\} = 0,
\end{aligned}$$

where we used (6.7) for the last inequality. For this reason and since $\bar{h} \in \mathcal{C}_{\bar{u}}$, (SSC) implies that $\bar{h} = 0$. With an analogous argument as in step 4 of the proof of Theorem 4.1 in [7], the desired contradiction is finally obtained. □

Remark 6.1. In general, the above second-order analysis does not apply to more general objective functionals of the form

$$J(\nabla y, u) = \int_{\Omega_g} f(\nabla y)\, dx + \frac{\beta}{2} \int_{\Omega_s} u^2\, dx,$$

even if $\Psi_y(x) = f(\nabla y(x))$ considered as a mapping from Ω_g to $\mathbb{R}$ is sufficiently smooth for all fixed $y \in H^1(\Omega_g)$. For the differentiability of J and f, respectively, one would in general need $W^{1,\infty}$ regularity of the states, which cannot be expected in our case. In the case of (P), we benefit from the continuous Fréchet differentiability of $\Phi(v(x)) = |v(x) - z(x)|^2$ from $L^2(\Omega_g)^2$ to $L^1(\Omega_g)$.

Acknowledgments. The authors are grateful to Dr. J. Rehberg (WIAS Berlin) for some helpful discussions concerning Lemma 2.1. They also thank Prof. E. Casas (Universidad de Cantabria) and Prof. F. Tröltzsch (Technische Universität Berlin) for giving the access to [7].

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Pure Appl. Math. 65, Academic Press, New York, 1975.

[2] S. Agmon, A. Douglas, and L. Nirenberg, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions* I, Comm. Pure Appl. Math., 12 (1959), pp. 623–727.

[3] J.-J. Alibert and J.-P. Raymond, *Boundary control of semilinear elliptic equations with discontinuous leading coefficients and unbounded controls*, Numer. Funct. Anal. Optim., 3/4 (1997), pp. 235–250.

[4] H. W. Alt, *Lineare Funktionalanalysis*, Springer-Verlag, Berlin, 1999.

[5] J. Appell and P. P. Zabrejko, *Nonlinear Superposition Operators*, Cambridge University Press, Cambridge, UK, 1990.

[6] M. Bergounioux and K. Kunisch, *On the structure of the Lagrange multiplier for state-constrained optimal control problems*, Systems Control Lett., 48 (2002), pp. 16–176.

[7] E Casas, J. C. De Los Reyes, and F. Tröltzsch, *Sufficient second-order optimality conditions for semilinear control problems with pointwise state constraints*, SIAM J. Optim., 19 (2008), pp. 616–643.

[8] E. Casas, F. Tröltzsch, and A. Unger, *Second order sufficient optimality conditions for some state-constrained control problems of semilinear elliptic equations*, SIAM J. Control Optim., 38 (2000), pp. 1369–1391.

[9] E. Casas, *Control of an elliptic problem with pointwise state constraints*, SIAM J. Control Optim., 24 (1986), pp. 1309–1318.

[10] E. Casas, *Boundary control of semilinear elliptic equations with pointwise state constraints*, SIAM J. Control Optim., 31 (1993), pp. 993–1006.

[11] J. Elschner, J. Rehberg, and G. Schmidt, *Optimal regularity for elliptic transmission problems including C^1 interfaces*, Interfaces Free Bound., 2 (2007), pp. 233–252.

[12] K. Gröger, *A $W^{1,p}$-estimate for solutions to mixed boundary value problems for second order elliptic differential equations*, Math. Ann., 283 (1989), pp. 679–687.

[13] O. Klein, P. Philip, and J. Sprekels, *Modeling and simulation of sublimation growth of SiC bulk single crystals*, Interfaces Free Bound., 6 (2004), pp. 295–314.

[14] M. Laitinen and T. Tiihonen, *Conductive-radiative heat transfer in grey materials*, Quart. Appl. Math., 59 (2001), pp. 737–768.

[15] C. Meyer, P. Philip, and F. Tröltzsch, *Optimal control of a semilinear PDE with nonlocal radiation interface conditions*, SIAM J. Control Optim., 45 (2006), pp. 699–721.

[16] C. Meyer and P. Philip, *Optimizing the temperature profile during sublimation growth of sic single crystals: Control of heating power, frequency, and coil position*, Crystal Growth & Design, 5 (2005), pp. 1145–1156.

[17] C. Meyer, *An SQP active set method for a semilinear optimal control problem with nonlocal radiation interface conditions*, in Control of Coupled Partial Differential Equations, K. Kunisch, G. Leugering, F. Tröltzsch, and J. Sprekels, eds., Internat. Ser. Numer. Math. 155, Birkhäuser Basel, 2007, pp. 217–248.

[18] C. Meyer, *Second-order sufficient optimality conditions for a semilinear optimal control problem with nonlocal radiation interface conditions*, ESAIM Contol Optim. Calc. Var., 13 (2007), pp. 750–775.

[19] P. Philip, *Transient Numerical Simulation of Sublimation Growth of SiC Bulk Single Crystals. Modeling, Finite Volume Method, Results*, Ph.D. thesis, Department of Mathematics, Humboldt University of Berlin, Berlin, 2003.

[20] J. Serrin, *Pathological solutions of elliptic differential equations*, Ann. Scuola Norm. Sup. Pisa (3), 18 (1964), pp. 385–387.

[21] T. Tiihonen, *A nonlocal problem arising from heat radiation on non-convex surfaces*, European J. Appl. Math., 8 (1997), pp. 403–416.

[22] E. Zeidler, *Nonlinear Functional Analysis and Its Application* II/*B: Nonlinear Monotone Operators*, Springer, New York, 1990.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 2, pp. 756–762

© 2009 Society for Industrial and Applied Mathematics

OPTIMALITY OF AN (s, S) POLICY WITH COMPOUND POISSON AND DIFFUSION DEMANDS: A QUASI-VARIATIONAL INEQUALITIES APPROACH[*]

LAKDERE BENKHEROUF[†] AND ALAIN BENSOUSSAN[‡]

Abstract. This paper revisits the paper of Bensoussan, Liu, and Sethi where they proposed a single item continuous-time inventory model where demand is a mixture of a diffusion process and a compound Poisson process. They showed that in a continuous review setting an (s, S) policy is optimal when the jump sizes are exponentially distributed. However, the case where the jump sizes are general was not solved completely. This paper solves this case.

Key words. inventory control, impulse control, quasi-variational inequality

AMS subject classifications. 49J15, 49J40, 65C30

DOI. 10.1137/080715883

1. Introduction.

In a recent paper Bensoussan, Liu, and Sethi (BLS) [1] proposed a single item continuous-time continuous state inventory model where demand is a mixture of a diffusion process and a compound Poisson process. They showed that in a continuous review setting an (s, S) policy is optimal when the jump sizes are exponentially distributed. However, the case where the demand is a combination of a diffusion and a general compound Poisson process was not completely resolved. This short note brings closure to this problem. The presentation builds on the paper of BLS and uses mostly the same notation.

Let $\xi_i \geq 0$, $i = 1, 2, \ldots$, be a sequence of independently and identically nonnegative random variables distributed having density $\mu(.)$ and finite mean. Also, let $N(t)$ be a right-continuous process with $N(0) = 0$ and intensity $\lambda > 0$ defined by

$$N(t) = \sum_i \xi_i \chi_{[0,t]}(\tau_i), \qquad t \geq 0, \tag{1.1}$$

where τ_i is the sequence of jump times and χ_A denotes the indicator function defined by $\chi_A(x) = 1$ if $x \in A$ and 0 otherwise.

We assume that the cumulative demand $y(t)$ on the interval $[0, t]$ is a stochastic process given by

$$y(t) = Dt + \sigma w(t) + N(t), \tag{1.2}$$

where $D \geq 0$ is constant, $w(t)$ is the standard Brownian motion with $w(0) = 0$, $\sigma \geq 0$, and $N(t)$ is defined in (1.1). The processes $w(t)$, $N(t)$, and ξ_i are all independent.

Let f be a real-valued function representing the holding and shortage cost with $f(0) = 0$ and $f(x) > 0$ for $x \neq 0$. The cost $c(x)$ of ordering an amount x is given by

$$c(x) = \begin{cases} K + cx, & x > 0, \\ 0, & x = 0, \end{cases} \tag{1.3}$$

[*] Received by the editors February 17, 2008; accepted for publication (in revised form) October 17, 2008; published electronically February 20, 2009.
http://www.siam.org/journals/sicon/48-2/71588.html

[†] Department of Statistics and Operations Research, College of Science, Kuwait University, P.O. Box 5969, Safat 13060, Kuwait (lakdereb@yahoo.com).

[‡] School of Management, University of Texas at Dallas, P.O. Box 830688, SM 30, Richardson, TX 75083-0688 (alain.bensoussan@utdallas.edu).

where $K > 0$ is the fixed setup cost of ordering and c is the unit cost of the item. Costs are assumed to be additive and exponentially discounted at a rate $\rho > 0$.

An admissible replenishment policy consists of a sequence (θ_i, u_i), $i = 1, \ldots,$ where θ_i represents the ith time of ordering and $u_i > 0$ represents the quantity ordered at time θ_i. Write

$$\begin{aligned} \mathcal{U}_n &= \{(\theta_i, u_i),\ i = 1, \ldots n\}, \\ \mathcal{U}_\infty &= \lim_{n\to\infty} \mathcal{U}_n = \mathcal{U}. \end{aligned}$$

Let x_t denote the level of stock at time t and $\mathcal{F}_n = \sigma\{x_s, s \le t\}$ be the σ-algebra generated by the history of the inventory level up to time t. Assume that, for each $n \in \mathbb{N}, \mathcal{U}_n$ is $\mathcal{F}_n$-measurable. Then, for a given initial inventory level x and an ordering policy $\mathcal{U}$, the discounted cost is defined by

$$F(x,\mathcal{U}) = E_{\mathcal{U}} \left\{ \int_0^\infty f(x(t)) e^{-\rho t} dt + \sum_{i=1}^n c(u_i) e^{-\rho\theta_i} \right\}, \tag{1.4}$$

where the expectation is taken with respect to all possible realizations of the process x_t under policy $\mathcal{U}$. Set

$$F(x) = \inf_{\mathcal{U}} F(x,\mathcal{U}).$$

The objective is to find an admissible policy $\mathcal{U}^*$ such that $F(x,\mathcal{U}^*) = F(x)$. Let

$$\begin{aligned} G(x) &:= F(x) + cx, \\ g(x) &:= f(x) + c\rho x, \end{aligned}$$

and introduce the operators A, B, and M defined for a function ϕ by

$$\begin{cases} (A\phi)(x) = -\dfrac{\sigma^2}{2}\phi''(x) + D\phi'(x), \\ (B\phi)(x) = \lambda \displaystyle\int_0^\infty (\phi(x-\xi) - \phi(x))\mu(\xi) d\xi, \\ (M\phi)(x) = K + \inf\limits_{u\ge 0} \{\phi(x+u)\}. \end{cases}$$

Write

$$h(x) := g(x) + cD + c\lambda\bar{\xi}, \tag{1.5}$$

where

$$\bar{\xi} = \int_0^\infty \xi\mu(\xi)d\xi < \infty.$$

Note that the operator M defined above differs from that considered in BLS. In BLSs quasi-variational inequality (QVI) the operators act on the function F, while in the present paper the operators act on G. In any case, the two formulations are equivalent. Finding the optimal policy U^* reduces to the problem of finding the solution of the following QVI problem:

$$\begin{cases} \Lambda G \le h, \\ G \le MG, \\ (\Lambda G - h)(G - MG) = 0, \end{cases} \tag{1.6}$$

where

$$(\Lambda\phi)(x) = (A\phi)(x) - (B\phi)(x) + \rho\phi(x).$$

The following lists a number of assumptions that are found in BLS.

(A1) The function f is non-negative and piecewise differentiable with $f(0) = 0$. Also, f has polynomial growth rate.

(A2) There exists a $\bar{\pi}$ such that

$$\int_0^\infty e^{-\pi\xi}\mu(\xi)d\xi < \infty \text{ for all } \pi > \bar{\pi}.$$

(A3) There exists a number a such that g is decreasing and convex on $(-\infty, a)$ and increasing on (a, ∞). Furthermore, there exist $c_0 > 0$ and $\eta \geq a$ such that $g'(x) \geq c_0$ for all $x \geq \eta$.

It is also implicit in BLS that $\int_0^\infty e^{-\bar{\pi}\xi}\mu(\xi)d\xi = \infty$.

Note that the function h defined in (1.5) inherits from g all of the assumptions made in (A3).

BLS showed that under assumptions (A1)–(A3) and the jump sizes exponentially distributed the optimal solution of the QVI (1.6) exists and is described by a pair (s, S) which corresponds to the well-known base stock policy in inventory control. However, the case where the demand is a combination of a diffusion and a general compound Poisson process was not completely resolved. The problem with the general case is resolved in the next section.

2. Solution of the QVI problem. BLS in their quest to find the solution of the QVI described in (1.6) considered the integro differential equation

(2.1)
$$-\frac{\sigma^2}{2}G''(x) + DG'(x) + (\rho + \lambda)G(x) - \lambda\int_0^\infty G(x-\xi)\mu(\xi)d\xi = h(x), \quad x > s,$$
$$G(x) = G(s), \quad x \leq s,$$

for some parameter $s < a$. They showed, using Laplace transform techniques, that under assumptions (A1)–(A3) a solution G_s of (2.1) exists. Also, the function G_s possesses the following properties:

(2.2) G_s is continuously differentiable on $(-\infty, \infty)$;

(2.3) G_s is constant on $(-\infty, s]$;

(2.4) G_s decreases on the interval $[s, a_0]$;

(2.5) $G_s(x) \to \infty$ as $x \to \infty$.

Here, $a_0 < a$ and is the solution of $Q(x) = 0$, where

$$Q(x) = \frac{2}{\sigma^2}\int_x^\infty e^{-\beta_2(y-x)}g'(y)dy,$$

and β_2 is some strictly positive parameter.

It follows from the properties of G_s (see BLS) that, for a given $s < a_0$, there exists an $S(s) > a_0$ where the function G_s reaches its minimum (Theorem 5.3 of BLS). It was also shown (Theorem 6.1 of BLS) that there exists a unique $s < a_0$ such that

$$G_s(s) = K + G_s(S(s)). \tag{2.6}$$

Now, let $(s, S(s))$ be a solution of (2.6). Note that s is unique but $S\,(s)$ may not be unique. Our proof of optimality relies on the concept related to non-K-decreasing functions which may be found in Porteus [2, p. 137].

Definition 1. *A function $f : \mathbb{R} \to \mathbb{R}$ is non-K-decreasing if*

$$f(x) \leq K + f(y) \quad \textit{for all } x \leq y.$$

Note that the concept of non-K-decreasing is weaker than the concept of an K-convexity which is a standard tool for showing optimality of an (s, S) policy; see Porteus [2] for more details. It is also immediate to deduce that a function f is non-K-decreasing if and only if for all $x \in \mathbb{R}$,

$$f(x) \leq K + \inf_{y \geq x} f(y).$$

Our objective is to show that the function G_s is non-K-decreasing. Note from properties (2.3)–(2.5) that G_s is constant on $(-\infty, s]$, then decreases at least up to a_0, reaches it minimum at some $S(s)$, and eventually goes to ∞ as $x \to \infty$. As x varies from $-\infty$ to ∞, non-K-decreasing means that the value $G_s(x)$ cannot have a drop bigger than K beyond $S(s)$.

Theorem 1. *Let G_s be the solution of* (2.1)*; then*

$$G_s(x) \leq K + \min_{y \geq x} G_s(y).$$

Proof. Define

$$B_s(x) = G_s(x) - \min_{y \geq x} G_s(y); \tag{2.7}$$

then the statement of the theorem is equivalent to

$$B_s(x) \leq K.$$

(i) If $x \leq s$, then (2.1) and (2.6) lead to $G_s(x) = G_s(s) = K + \min_y G_s(y) = K + \min_{y \geq x} G_s(y)$ or $B_s(x) = K$. Therefore, the theorem is true in this case.

(ii) If $s \leq x \leq a_0$, then Property (2.4) shows that

$$G_s(x) < G_s(s) = K + \min_{y} G_s(y) = K + \min_{y \geq x} G_s(y).$$

This leads to the required result in this case.

(iii) If $x > a_0$ and $B_s(x) > K$, then we shall show that there exist points x', x_1, x_2, and x_3 with $x_1 < x' \leq x_2 < x_3$. Properties of the function G_s at these points is then shown to lead to a contradiction; see Figure 1.

Define

$$x' = \inf\,\{x > a_0,\, B_s(x) > K\}.$$

It follows that $B_s(x') = K$ and $B_s(x) < K$ for $s < x < x'$.

Put $\gamma = \min_{x \geq x'} G_s(x)$ and

$$x_3 = \min\,\{x \geq x' : G_s(x) = \gamma\}. \tag{2.8}$$

The point x_3 exists by (2.2) and (2.5). It is also a local minimum, and consequently $G_s'(x_3) = 0$, $G_s''(x_3) \geq 0$, and

$$G_s(x') - G_s(x_3) = K. \tag{2.9}$$

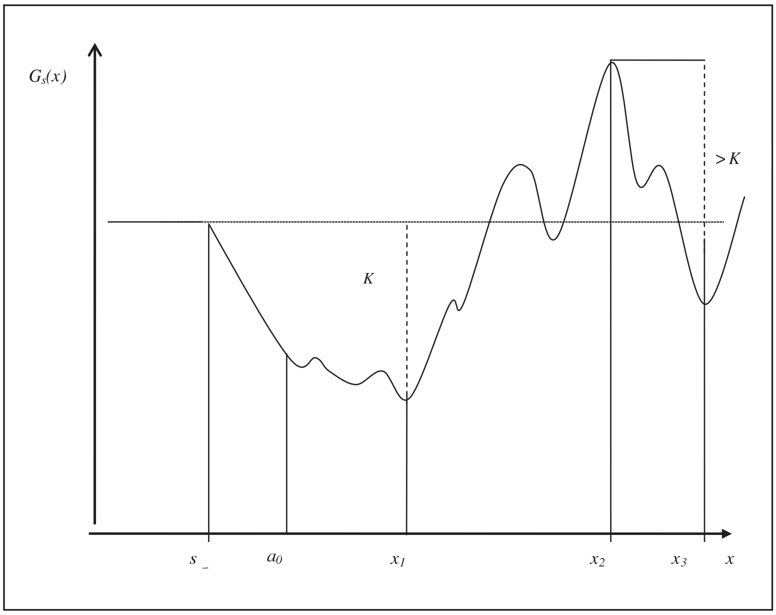


FIG. 1. *Typical inventory behavior.*

Put $\Gamma = \max_{x \le x_3} G_s(x)$ and

$$x_2 = \min\,\{x \le x_3 : G_s(x) = \Gamma\}. \tag{2.10}$$

The point x_2 exists since G_s is continuous. It is a local maximum with $G_s'(x_2) = 0$ and $G_s''(x_2) \le 0$. Further, $x' \le x_2$. Indeed, if $x_2 < x'$, it follows from the definition of x' that $G_s(x_2) - G_s(x_3) < K$ or $G_s(x_2) < G_s(x_3) + K = G_s(x')$, which is in contradiction with the definition of x_2.

Note that (2.9) and (2.10) give

$$K \le G_s(x_2) - G_s(x_3). \tag{2.11}$$

Let

$$\zeta = \min_{x \le x'} G_s(x)$$

and

$$x_1 = \min\{x \le x',\ G_s(x) = \zeta\}. \tag{2.12}$$

The point x_1 exists by (2.2). Also, x_1 is a local minimum with $G_s'(x_1) = 0$ and $G_s''(x_1) \ge 0$.

Now, we have completed the construction of the required points with $a_0 < x_1 < x' \le x_2 < x_3$. A crucial property is next sought before we move to the contradiction argument.

We claim that G_s possesses the following property:

$$G_s(x) - G_s(y) \le G_s(x_2) - G_s(x_3) \text{ for } x \le y \le x_3. \tag{2.13}$$

If $x < x'$, then

$$G_s(x) - G_s(y) \leq G_s(x) - \min_{\eta \geq x} G_s(\eta) < K \leq G_s(x_2) - G_s(x_3).$$

The last inequality follows from (2.11).

If $x' \leq x \leq y \leq x_3$, the result is immediate from the definitions of x_2 and x_3.

Now, we are ready for the contradiction argument. We examine two possible cases relating to the position of x_2 with respect to a.

(v) If $a_0 < x_2 \leq a$, then (2.1) with the definitions of x_1 and x_2 gives

$$\rho G_s(x_1) = h(x_1) - \lambda \int_0^\infty (G_s(x_1) - G_s(x_1 - \xi))\mu(\xi)d\xi + \frac{\sigma^2}{2} G_s''(x_1), \tag{2.14}$$

$$\rho G_s(x_2) = h(x_2) - \lambda \int_0^\infty (G_s(x_2) - G_s(x_2 - \xi))\mu(\xi)d\xi + \frac{\sigma^2}{2} G_s''(x_2). \tag{2.15}$$

Note by (2.12) that $G_s(x_1) - G_s(x_1 - \xi) \leq 0$ for all $\xi \geq 0$ and by (2.10) that $G_s(x_2) - G_s(x_2 - \xi) \geq 0$ for all $\xi \geq 0$. Also, $G_s''(x_1) \geq 0$ and $G_s''(x_2) \leq 0$. Furthermore, h is decreasing on $(-\infty, a]$ by assumption (A3). It follows that $\rho G_s(x_2) \leq \rho G_s(x_1)$, which leads to a contradiction since $G_s(x_2) > G_s(x_1)$. Hence, $B_s(x) \leq K$ in this case.

(vi) If $x_2 > a$, then (2.1) gives

$$(\rho + \lambda) G_s(x_2) = h(x_2) + \lambda \int_0^\infty G_s(x_2 - \xi)\mu(\xi)d\xi + \frac{\sigma^2}{2} G_s''(x_2), \tag{2.16}$$

$$(\rho + \lambda) G_s(x_3) = h(x_3) + \lambda \int_0^\infty G_s(x_3 - \xi)\mu(\xi)d\xi + \frac{\sigma^2}{2} G_s''(x_3). \tag{2.17}$$

The properties of G_s at x_2 and x_3 imply that

$$(\rho + \lambda)(G_s(x_2) - G(x_3)) \leq h(x_2) - h(x_3) + \lambda \int_0^\infty (G_s(x_2 - \xi) - G_s(x_3 - \xi))\mu(\xi)d\xi.$$

Also, by property (2.13) we have for $\xi \geq 0$,

$$G_s(x_2 - \xi) - G_s(x_3 - \xi) \leq G_s(x_2) - G_s(x_3)$$

since $x_2 - \xi \leq x_3 - \xi \leq x_3$. Therefore, $\rho(G_s(x_2) - G(x_3)) \leq h(x_2) - h(x_3) < 0$ since h is increasing on $[a, \infty)$ and μ is a density function. But (2.11) gives $G_s(x_2) - G_s(x_3) \geq K > 0$. This leads to a contradiction. This completes the proof. □

Define $S(s)$ to be the smallest S which solves

$$G_s(s) = K + G_s(S). \tag{2.18}$$

Theorem 2. *Let* $(s, S(s))$ *be a solution of* (2.18)*; then this solution solves the QVI* (1.6).

Proof. We will show that $\Lambda G \leq h$ on $(-\infty, \infty)$ with equality on $[s, \infty)$ and that $G \leq MG$ on $(-\infty, \infty)$ with equality on $(-\infty, s]$. By construction we have $\Lambda G = h$ on $[s, \infty)$. Furthermore, on $(-\infty, s]$, $\Lambda G \leq h$ is equivalent to $h(s) \leq h(x)$, which is true since h is decreasing on $(-\infty, a]$ and $s < a_0 < a$. Now, turning to the operator M, we have

$$MG(x) = \begin{cases} K + G_s(S(s)) & \text{if } x \leq S(s), \\ K + \inf\limits_{u \geq 0} G_s(x + u) & \text{if } x > S(s). \end{cases}$$

If $x \leq s$, the result $MG(x) = G_s(s) = K + G_s(S(s))$. If $s < x \leq S(s)$, then $G \leq MG$ is equivalent to $G_s(x) \leq K + G_s(S(s))$, which is true by Theorem 1. Finally, if $x > S(s)$, we have

$$MG(x) = K + \inf_{u \geq 0} G_s(x+u) = K + \inf_{y \geq x} G_s(y).$$

Again, $G \leq MG$ is equivalent to $G(x) - \inf_{y \geq x} G_s(y) \leq K$, which follows from Theorem 1 leading to $G \leq MG$ on $[S(s), \infty)$. This completes the proof. ▯

In this paper we have shown that an (s, S) policy is optimal for an inventory model in which demand is a combination of a diffusion and a general compound process. This solves the open problem that was left in BLS.

Acknowledgment. Lakdere Benkherouf would like to thank Dr. Michael Johnson for useful comments on an earlier version of the paper.

REFERENCES

[1] A. Bensoussan, R.H. Liu, and S.P. Sethi, *Optimality of an (s, S) policy with compound Poisson and diffusion demands: A quasi-variational inequalities approach*, SIAM J. Control Optim., 44 (2005), pp. 1650–1676.

[2] E.L. Porteus, *Foundations of Stochastic Inventory Theory*, Stanford University Press, Palo Alto, CA, 2002.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 2, pp. 763–786

© 2009 Society for Industrial and Applied Mathematics

STRONG STABILITY OF NEUTRAL EQUATIONS WITH AN ARBITRARY DELAY DEPENDENCY STRUCTURE[*]

WIM MICHIELS[†], TOMÁŠ VYHLÍDAL[‡], PAVEL ZÍTEK[‡], HENK NIJMEIJER[§], AND DIDIER HENRION[¶]

Abstract. The stability theory for linear neutral equations subjected to delay perturbations is addressed. It is assumed that the delays cannot necessarily vary independently of each other, but depend on a possibly smaller number of independent parameters. As a main result, necessary and sufficient conditions for strong stability are derived along with bounds on the spectrum, which take into account the precise dependency structure of the delays. In the derivation of the stability theory, results from realization theory and determinantal representations of multivariable polynomials play an important role. The observations and results obtained in the paper are first illustrated and validated with a numerical example. Next, the effects of small feedback delays on the stability of a boundary controlled hyperbolic partial differential equation and of a control system involving state derivative feedback are analyzed.

Key words. neutral system, strong stability, spectral theory

AMS subject classifications. 93D09, 93D20, 93C23

DOI. 10.1137/080724940

Notation.

$\mathbb{C}$	set of complex numbers
$\mathbb{C}^-,\ \mathbb{C}^+$	open left half plane, open right half plane
i	imaginary identity
$\mathbb{N}$	set of natural numbers, including zero
$\mathbb{R}$	set of real numbers
$\mathbb{R}^+$	$\{r \in \mathbb{R} :\ r \geq 0\}$
$\mathbb{R}_0^+$	$\mathbb{R}^+ \setminus \{0\}$
$e_k \in \mathbb{N}^m$	kth unit vector in $\mathbb{N}^m$
$\Re(\lambda), \Im(\lambda), \lvert\lambda\rvert,\ \lambda \in \mathbb{C}$	real part, imaginary part, and modulus of λ
$\vec{r} \in \mathbb{R}^m, \vec{n} \in \mathbb{N}^m, \dots$	short notation for $(r_1, \dots, r_m),\ (n_1, \dots, n_m), \dots$
$r_\sigma(A)$	spectral radius of operator (or matrix) A
$r_e(A)$	radius of the essential spectrum of operator A
$\sigma(A)$	spectrum of operator (or matrix) A
$\sigma_e(A)$	essential spectrum of operator (or matrix) A

[*] Received by the editors May 22, 2008; accepted for publication (in revised form) October 29, 2008; published electronically February 20, 2009. This research was supported by the Belgian Programme on Interuniversity Poles of Attraction, initiated by the Belgian State, Prime Ministers Office for Science, Technology and Culture, and by OPTEC, the Center of Excellence on Optimization in Engineering of the K.U.Leuven. This research was also supported by the Ministry of Education of the Czech Republic under Research Program MSM6840770038 and Research Program 1M0567, and by the Grant Agency of the Czech Republic under Project 102/06/0652.

http://www.siam.org/journals/sicon/48-2/72494.html

[†] Department of Computer Science, Katholieke Universiteit Leuven, Belgium (Wim.Michiels@cs.kuleuven.be). This work was partly done while the first author was with the Department of Mechanical Engineering at the Eindhoven University of Technology.

[‡] Centre for Applied Cybernetics, Department of Instrumentation and Control Eng., Faculty of Mechanical Eng., Czech Technical University in Prague, Czech Republic (Tomas.Vyhlidal@fs.cvut.cz, Pavel.Zitek@fs.cvut.cz).

[§] Department of Mechanical Engineering, Eindhoven University of Technology, The Netherlands (h.nijmeijer@tue.nl).

[¶] LAAS-CNRS, University of Toulouse, France (henrion@laas.fr).

$\mathrm{sign}(x),\ x \in \mathbb{R}$	$\mathrm{sign}(x) = \begin{cases} 1, & x \geq 0 \\ -1, & x < 0 \end{cases}$
$\mathbb{Z}$	set of integer numbers
$\alpha(A)$	spectral abscissa of operator (or matrix) A, $\alpha(A) := \sup\{\Re(\lambda) :\ \lambda \in \mathbb{C} \text{ and } \lambda \in \sigma(A)\}$
$\|\vec{a}\|,\ \vec{a} \in \mathbb{R}^m$	Euclidean norm of $\vec{a}$, $\|\vec{a}\| := \sqrt{\sum_{k=1}^m a_k^2}$
$\vec{a} \cdot \vec{b},\ \vec{a}, \vec{b} \in \mathbb{R}^m$	Euclidean inner product of $\vec{a}$ and $\vec{b}$, $\vec{a} \cdot \vec{b} := \sum_{k=1}^m a_k b_k$

1. Introduction. Many engineering systems can be modeled by delay differential equations of neutral type, for instance, lossless transmission lines [17] and partial element equivalent circuits [4] in electrical engineering, and combustion systems [26] and controlled constrained manipulators [27] in mechanical engineering. Equations of neutral type also arise in boundary-controlled hyperbolic partial differential equations subjected to small feedback delays [24, 6] and in implementation schemes of predictive controllers for time-delay systems [7, 25]. In this paper we discuss stability properties of the linear neutral equation

$$\dot{x}(t) + \sum_{k=1}^{p_1} H_k \dot{x}(t - \tau_k) = A_0 x(t) + \sum_{k=1}^{p_2} A_k x(t - \upsilon_k), \tag{1.1}$$

where $x(t) \in \mathbb{R}^n$ is the state variable at time t, $\vec{\tau} := (\tau_1, \ldots, \tau_{p_1}) \in (\mathbb{R}_0^+)^{p_1}$ and $\vec{\upsilon} := (\upsilon_1, \ldots, \upsilon_{p_2}) \in (\mathbb{R}_0^+)^{p_2}$ are time-delays, and H_k and A_k are real matrices.

An important aspect in the stability theory of neutral equations is the possible fragility of stability, in the sense that the asymptotic stability of the null solution of (1.1) may be sensitive to *arbitrarily small perturbations* of the delays $\vec{\tau}$; see, e.g., [12, 21, 24, 18] and the references therein. This has led to the introduction of the notion of strong stability in [11, 13, 14], which explicitly takes into account the effect of small delay perturbations. In [13] a necessary and sufficient condition for the strong stability of the null solution of (1.1) is described for the special case where the delays $(\tau_1, \ldots, \tau_m)$ can vary independently of each other (see also [9]), and in [23] some related spectral properties are discussed, though the focus lies on a stabilization procedure for systems with an external input. Note that robustness against delay perturbations is of primary interest in control problems, as parametric uncertainty and feedback delays are inherent features of control systems.

In the existing literature on the stability of neutral equations, subjected to delay perturbations, the delays, τ_k, $1 \leq k \leq p_1$, in (1.1) are almost exclusively assumed to be either mutually independent or commensurate (all multiples of the same parameter); an exception is formed by [28] where a problem with three delays depending on two independent parameters is analyzed. In this paper we study the dependence of the stability properties of (1.1) on the delay parameters, under the assumption that the delays τ_k, $1 \leq k \leq p_1$, are linear functions of $m \geq 1$ "independent" parameters $\vec{r} = (r_1, \ldots, r_m) \in (\mathbb{R}_0^+)^m$, as described by the following relation:

$$\tau_k = \vec{\gamma}_k \cdot \vec{r}, \quad k = 1, \ldots, p_1, \tag{1.2}$$

with

$$\vec{\gamma}_k := (\gamma_{k,1}, \ldots, \gamma_{k,m}) \in \mathbb{N}^m \setminus \{\vec{0}\}, \quad k = 1, \ldots, p_1.$$

Note that the cases of mutually independent delays, respectively, commensurate delays, appear in this framework as extreme cases ($m = n$ and $\vec{\gamma}_k = \vec{e}_k$, $k = 1, \ldots, p_1$, respectively, $m = 1$). The problem studied in [28] corresponds to the relation $(\tau_1, \tau_2, \tau_3) = (r_1, r_2, r_1 + r_2)$, which is also of the form (1.2).

There are several main reasons why it is important to develop a stability theory where any delay dependency structure of the form (1.2) can be taken into account explicitly. First, real systems might give rise to a model of the form (1.1) exhibiting a delay dependency caused by physical or other interactions in the system's dynamics. This is explained with a lossless transmission line example in Chapter 9.6 of [11], where it is shown that a parallel transmission line which consists of a current source, two resistors, and a capacitor gives rise to a system of a neutral type with three delays in the difference part, which are integer combinations of two physical parameters. In [6, 19, 24] boundary-controlled partial differential equations are described that lead to a closed-loop system of neutral type, where the delays in the model are particular linear combinations of (physical) feedback delays and delays induced by propagation phenomena. In [31, 32] the robustness against small feedback delays of linear systems controlled with state derivative feedback is addressed, motivated by vibration control applications. There, the closed-loop system can again be written in the form (1.1), where the delays τ_k are combinations of actuator and sensor delays in input and output channels. All these applications give rise to a (nonextreme case of a) delay dependency of the form (1.2). Second, the precise dependency of the delays has a *major* influence on the stability robustness. For instance, we shall illustrate that the asymptotic stability of (1.1) may be destroyed by arbitrarily small perturbations of the delays τ_k, $1 \leq k \leq p_1$, if these perturbations can be chosen independently of each other, but it may be robust against small perturbations if the (perturbed) delays are restricted by a relation like (1.2). Third, the analysis for an arbitrary delay dependence of the form (1.2) is much more complex than the analysis of the special cases available in the literature (e.g., fully independent delays in [23]), where the derivation of the results heavily relies on specific properties induced by the special case. In this discussion it is worthwhile to note that no assumptions need to be made on the interdependency of the delays $\vec{\nu}$, because, as we shall see, this interdependency does not affect the stability robustness with respect to (w.r.t.) small delay perturbations, unlike the interdependency of the delays $\vec{r}$.

While the general aim of the paper is to develop a stability theory for neutral equations with dependent delays subjected to delay perturbations, the emphasis is on the derivation of explicit strong stability criteria and on related spectral properties. As we shall see, only in specific situations, where severe restrictions are put on the dependency structure, can the criteria available in the literature for independent delays be directly generalized, though the derivation is more complicated. To obtain a *general* solution and, in this way, complete the theory, some type of intermediate lifting step may be necessary, where a delay difference equation with dependent delays is transformed into an equation with independent delays with the same spectral properties. The main step will boil down to the representation of a multivariable polynomial as the determinant of a pencil. Such a representation will follow from arguments of realization theory, more precisely, from the construction of lower fractional representations (LFRs). See, for instance, [33] and the manual of the LFR toolbox [20] for an introduction.

Finally, we note that the strong stability criteria developed in this paper are also important in the context of stabilization and control of neutral systems. If the null solution of the associated difference equation is strongly stable, then the unstable manifold is finite-dimensional and remains so in the presence of delay perturbations.

This opens the possibilities of using controllers which act only on that manifold (see, e.g., [29]) or which are based on shifting or assigning a finite number of eigenvalues as [24]. On the contrary, if the difference equation is not strongly stable, then the closed-loop system lacks robustness against small delay perturbations. This may happen even if the application of the control law involves a noncompact perturbation of the solution operator and, thus, directly affects the difference equation; see [13] for an illustration.

The structure of the paper is as follows: in section 2 some basic notions and results on neutral equations are recalled, in support of the subsequent sections. In section 3 the spectral properties of the neutral equation (1.1)–(1.2) and of the associated delay difference equation are addressed, with the emphasis on stability properties and the sensitivity of stability w.r.t. delay perturbations. The main results are presented in section 4, where computational expressions are presented that lead to explicit strong stability conditions. Section 5 is devoted to applications and illustrations. Section 6 contains the conclusions.

2. Preliminaries.

The initial condition for the neutral system (1.1)–(1.2) is a function segment $\varphi \in \mathcal{C}([-\bar{\tau},\ 0], \mathbb{R}^n)$, where $\bar{\tau} = \max_{k\in\{1,\dots,p_1\}} \tau_k$ and $\mathcal{C}([-\bar{\tau},\ 0],\ \mathbb{R}^n)$ is the Banach space of continuous functions mapping the interval $[-\bar{\tau},\ 0]$ into $\mathbb{R}^n$ and equipped with the supremum-norm. The fact that the map $\mathcal{D}:\ \mathcal{C}([-\bar{\tau},\ 0], \mathbb{R}^n) \to \mathbb{R}^n$, defined by

$$\mathcal{D}(\varphi) = \varphi(0) + \sum_{k=1}^{p_1} H_k \varphi(-\tau_k),$$

is atomic at zero guarantees existence and uniqueness of solutions of (1.1). Let $x(\varphi) : t \in [-\bar{\tau},\ \infty) \to x(\varphi)(t) \in \mathbb{R}^n$ be the unique forward solution with initial condition $\varphi \in \mathcal{C}([-\bar{\tau},\ 0], \mathbb{R}^n)$, i.e., $x(\varphi)(\theta) = \varphi(\theta)$ for all $\theta \in [-\bar{\tau},\ 0]$. Then the state at time t is given by the function segment $x_t(\varphi) \in \mathcal{C}([-\bar{\tau},\ 0],\ \mathbb{R}^n)$ defined as $x_t(\varphi)(\theta) = x(\varphi)(t+\theta),\ \theta \in [-\bar{\tau},\ 0]$. Denote by $\mathcal{T}(t; \vec{r}, \vec{v})$ the solution operator, mapping initial data onto the state at time t, i.e.,

$$(\mathcal{T}(t;\ \vec{r}, \vec{v})\varphi)(\theta) = x_t(\varphi)(\theta) = x(\varphi)(t+\theta), \quad \theta \in [-\bar{\tau},\ 0]. \tag{2.1}$$

This is a strongly continuous semigroup. The associated delay difference equation of (1.1) is given by

$$z(t) + \sum_{k=1}^{p_1} H_k z(t - \vec{\gamma}_k \cdot \vec{r}) = 0. \tag{2.2}$$

For any initial condition $\varphi \in \mathcal{C}_D([-\bar{\tau},\ 0], \mathbb{R}^n)$, where

$$\mathcal{C}_D([-\bar{\tau},\ 0], \mathbb{R}^n) = \left\{\varphi \in \mathcal{C}([-\bar{\tau},\ 0], \mathbb{R}^n) :\ \mathcal{D}(\varphi) = 0\right\},$$

a solution $z(\varphi)(t)$ of (2.2) is uniquely defined and satisfies $z_t(\phi) \in \mathcal{C}_D([-\bar{\tau},\ 0], \mathbb{R}^n)$ for all $t \geq 0$. Let $\mathcal{T}_D(t;\ \vec{r})$ be the corresponding solution operator.

The asymptotic behavior of the solutions and, thus, the stability of the null solution of the neutral equation (1.1) is determined by the spectral radius $r_\sigma(\mathcal{T}(t;\ \vec{r}, \vec{v}))$, satisfying

$$\begin{aligned} & r_\sigma(\mathcal{T}(1;\ \vec{r}, \vec{v})) = \epsilon^{c_N(\vec{r}, \vec{v})}, \\ & c_N(\vec{r}, \vec{v}) = \sup\left\{\Re(\lambda) :\ \det\left(\Delta_N(\lambda;\ \vec{r}, \vec{v})\right) = 0\right\}, \end{aligned} \tag{2.3}$$

where the characteristic matrix Δ_N is given by

$$\Delta_N(\lambda;\ \vec{r},\vec{v}) = \left(\lambda\Delta_D(\lambda;\ \vec{r}) - A_0 - \sum_{k=1}^{p_2} A_k e^{-\lambda v_k}\right) \tag{2.4}$$

and

$$\Delta_D(\lambda;\ \vec{r}) = \left(I + \sum_{k=1}^{p_1} H_k e^{-\lambda \vec{\gamma}_k\cdot\vec{r}}\right).$$

For instance, the null solution is exponentially stable if and only if $r_\sigma(\mathcal{T}(1;\ \vec{r},\vec{v})) < 1$ or equivalently $c_N(\vec{r},\vec{v}) < 0$ [13, 12] (see [11] for an overview of stability definitions and their relation to spectral properties). In a similar way, the stability of the delay difference equation (2.2) is determined by the spectral radius

$$r_\sigma(\mathcal{T}_D(1;\ \vec{r})) = e^{c_D(\vec{r})}, \tag{2.5}$$

where

$$c_D(\vec{r}) = \begin{cases} -\infty, & \det(\Delta_D(\lambda;\ \vec{r})) \neq 0\ \forall\lambda\in\mathbb{C}, \\ \sup\left\{\Re(\lambda):\ \det\left(\Delta_D(\lambda;\ \vec{r})\right) = 0\right\}, & \text{otherwise.}\end{cases} \tag{2.6}$$

An important property in the stability analysis of neutral equations is the relation

$$r_e(\mathcal{T}(1;\ \vec{r},\vec{v})) = r_\sigma(\mathcal{T}_D(1;\ \vec{r})); \tag{2.7}$$

see, e.g., [11, 10]. From this follows the well-known result that a *necessary condition* for the exponential stability of the null solution of (1.1)–(1.2) is given by the exponential stability of the null solution of the delay difference equation (2.2).

In the remainder of the paper we will call the solutions of $\det(\Delta_N(\lambda;\ \vec{r},\vec{v})) = 0$ the characteristic roots of the neutral system (1.1). Analogously we will call the solutions of $\det(\Delta_D(\lambda;\ \vec{r})) = 0$ the characteristic roots of the delay difference equation (2.2).

3. Spectral properties.

We discuss some spectral properties of the neutral equation (1.1) which are important for the rest of the paper. In section 3.1–3.2 we make the implicit assumption that

$$\exists\lambda\in\mathbb{C}:\ \det\Delta_D(\lambda;\ \vec{r}) \not\equiv 1.$$

The degenerate case where this condition is not met will be treated separately in section 3.3.

3.1. Difference equation.

It is well known that the spectral radius (2.5), although continuous in the system matrices H_k, is *not* continuous in the *delays* $\vec{r}$ (see, e.g., [11, 13, 16, 23]), which carries over to (2.6). As a consequence, we are from a practical point of view led to the smallest upper bound on the real parts of the characteristic roots, which is "*insensitive*" to small delay changes.

Definition 3.1. *For $\vec{r}\in(\mathbb{R}_0^+)^m$, let $\bar{C}_D(\vec{r})\in\mathbb{R}$ be defined as*

$$\bar{C}_D(\vec{r}) = \lim_{\epsilon\to 0+} c_\epsilon(\vec{r}),$$

where

$$c_\epsilon(\vec{r}) = \sup\left\{c_D(\vec{r}+\delta\vec{r}):\ \delta\vec{r}\in\mathbb{R}^m \text{ and } \|\delta\vec{r}\|\leq\epsilon\right\}.$$

Clearly we have $\bar{C}_D(\vec{r}) \geq c_D(\vec{r})$, and the inequality can be *strict*, as shown in [23] and illustrated later on. We have the following results.

Proposition 3.2. *The following assertions hold:*

1. *the function*
$$\vec{r} \in (\mathbb{R}_0^+)^m \mapsto \bar{C}_D(\vec{r})$$
is continuous;
2. *for every* $\vec{r} \in (\mathbb{R}_0^+)^m$, *we have*[1]
$$\bar{C}_D(\vec{r}) = \max \left\{ c \in \mathbb{R} : \ \det\left(I + \sum_{k=1}^{p_1} H_k e^{-c\vec{\gamma}_k\cdot\vec{r}} e^{-i\vec{\gamma}_k\cdot\vec{\theta}} \right) = 0 \right. \tag{3.1}$$
$$\left. \textit{for some } \vec{\theta} \in [0,\ 2\pi]^m \right\};$$
3. $\bar{C}_D(\vec{r}) = c_D(\vec{r})$ *for rationally independent*[2] $\vec{r}$*;*
4. *for all* $\vec{r}_1, \vec{r}_2 \in (\mathbb{R}_0^+)^m$, *we have*
$$\operatorname{sign}\left(\bar{C}_D(\vec{r}_1)\right) = \operatorname{sign}\left(\bar{C}_D(\vec{r}_2)\right). \tag{3.2}$$

Proof. Assertions 1 and 3 are direct corollaries of Lemma 2.5 and Theorem 2.2 of [3]. Combining assertion 3 with Theorem 3.1 of [3] yields assertion 2. The proof of assertion 4 is by contradiction. If (3.2) is not satisfied, then by assertion 1 there exists a vector $\vec{s} \in (\mathbb{R}_0^+)^m$ for which $\bar{C}_D(\vec{s}) = 0$. This implies by (3.1) that $\bar{C}_D(\vec{r}) \geq 0$ for all $\vec{r} \in (\mathbb{R}_0^+)^m$ and we arrive at a contradiction. □

The property (3.2) leads us to the following definition.

Definition 3.3. *Let* $\Xi := \operatorname{sign}\left(\bar{C}_D(\vec{r})\right)$, $\vec{r} \in (\mathbb{R}_0^+)^m$.

A consequence of the *non*continuity of c_D w.r.t. $\vec{r}$ is that arbitrarily small perturbations on the delays may destroy stability of the delay difference equation. This phenomenon, which was illustrated in [24], has lead to the introduction of the concept of *strong stability* in [13]: we say that the null solution of (2.2) is strongly exponentially stable if it is exponentially stable and remains so when subjected to small variations in the delays $\vec{r}$. We state this more precisely in the following definition.

Definition 3.4. *The null solution of the delay difference equation* (2.2) *is strongly exponentially stable if there exists a number* $\hat{r} > 0$ *such that the null solution of*
$$z(t) + \sum_{k=1}^{p_1} H_k z(t - \vec{\gamma}_k \cdot (\vec{r} + \delta\vec{r})) = 0$$
is exponentially stable for all $\delta\vec{r} \in (\mathbb{R}^+)^m$ *satisfying* $\|\delta\vec{r}\| < \hat{r}$ *and* $r_k + \delta r_k > 0,\ 1 \leq k \leq m$.

The following condition follows from Proposition 3.2.

Proposition 3.5. *The null solution of* (2.2) *is strongly exponentially stable if and only if* $\Xi < 0$.

Proof. By definition the null solution of (2.2) is strongly exponentially stable if and only if $\bar{C}_D(\vec{r}) < 0$, which is equivalent to $\Xi < 0$. □

Remark 3.6. The condition of Proposition 3.5 does not depend on the particular value of $\vec{r} \in (\mathbb{R}_0^+)^m$, that is, strong exponential stability for one value of $\vec{r}$ implies strong exponential stability for all values of $\vec{r}$.

[1] The maximum in (3.1) is well defined because $\vec{\theta}$ belongs to a compact set.

[2] The m components of $\vec{r} = (r_1, \ldots, r_m)$ are rationally independent if and only if the conditions $\sum_{k=1}^m n_k r_k = 0$ and $n_k \in \mathbb{Z}$ imply $n_k = 0$ for all $k = 1, \ldots, m$. For instance, two delays r_1 and r_2 are rationally independent if their ratio is an irrational number.

3.2. Neutral equation. Following from (2.7), not only the delay difference equation (2.2) but also the neutral equation (1.1)–(1.2) have characteristic roots with real part *arbitrarily close* to $\bar{C}_D(\vec{r})$ for certain (arbitrarily small) perturbations on $\vec{r}$.

From the fact that the operator $\mathcal{T}(1;\ \vec{r}, \vec{\upsilon})$, defined in (2.1), has only a *point spectrum* in the set

$$\{\lambda \in \mathbb{C} :\ |\lambda| > r_e(\mathcal{T}(1;\ \vec{r}, \vec{\upsilon})) = r_\sigma(\mathcal{T}_D(1;\ \vec{r}))\}$$

(see [13]), it follows that all the characteristic roots of (1.1) in the half plane

$$\left\{\lambda \in \mathbb{C} :\ \Re(\lambda) \geq \bar{C}_D(\vec{r}) + \epsilon\right\},$$

where $\epsilon > 0$, lie in a *compact set* and that the number of these roots (multiplicity taken into account) is *finite*. Bounds on these roots can be obtained from the following lemma, whose proof can be found in Appendix A.

Lemma 3.7. *If* $\Delta_N(\lambda;\ \vec{r}, \vec{\upsilon}) = 0$ *and* $\Re(\lambda) > \bar{C}_D(\vec{r})$*, then*

$$|\lambda| \leq \max_{\vec{\theta} \in [0,\ 2\pi]^m} \left\| \left(I + \sum_{k=1}^{p_1} H_k e^{-\Re(\lambda)(\vec{\gamma}_k \cdot \vec{r})} e^{-i\vec{\gamma}_k \cdot \vec{\theta}} \right)^{-1} \right\| \left(\|A_0\| + \sum_{i=1}^{p_1} \|A_k\| e^{-\Re(\lambda)\upsilon_k} \right).$$

By combining the above results we arrive at the following result.

Proposition 3.8. *The function*

$$(\vec{r}, \vec{\upsilon}) \in (\mathbb{R}_0^+)^m \times (\mathbb{R}^+)^{p_2} \mapsto \max(\bar{C}_D(\vec{r}), c_N(\vec{r}, \vec{\upsilon}))$$

is continuous.

We refer to Appendix B for a detailed proof.

Proposition 3.8 is an important result, given that the function $(\vec{r}, \vec{\upsilon}) \in (\mathbb{R}_0^+)^m \times (\mathbb{R}^+)^{p_2} \mapsto c_N(\vec{r}, \vec{\upsilon})$ is not continuous, with discontinuities occurring at delay values where $c_N(\vec{r}, \vec{\upsilon}) < \bar{C}_D(\vec{r})$. Such situations do occur and will be illustrated in the first example of section 5.

Furthermore, if we define strong exponential stability for the neutral equation (1.1)–(1.2) analogously as for the associated delay difference equation, then we have the following definition.

Definition 3.9. *The null solution of the neutral equation* (1.1)–(1.2) *is strongly exponentially stable if there exists a number* $\hat{r} > 0$ *such that the null solution of*

$$\dot{x}(t) + \sum_{k=1}^{p_1} H_k \dot{x}(t - \vec{\gamma}_k \cdot (\vec{r} + \delta\vec{r})) = A_0 + \sum_{k=1}^{p_2} A_k x(t - (\upsilon_k + \delta\upsilon_k))$$

is exponentially stable for all $\delta\vec{r} \in (\mathbb{R}.^+)^m$ *and* $\delta\vec{\upsilon} \in (\mathbb{R}^+)^{p_2}$ *satisfying* $\|\delta\vec{r}\| < \hat{r},\ \|\delta\vec{\upsilon}\| < \hat{r}$ *and* $r_k + \delta r_k > 0,\ \nu_l + \delta\nu_l > 0,\ 1 \leq k \leq m,\ 1 \leq l \leq p_2$.

Then we get the following result.

Proposition 3.10. *The null solution of the neutral equation* (1.1) *is strongly exponentially stable if and only if* $c_D(\vec{r}, \vec{\nu}) < 0$ *and* $\Xi < 0$.

Remark 3.11. Proposition 3.10 implies that the interdependency of the delays $\vec{\upsilon}$, if any, does not affect the strong stability of the neutral equation (3.10), unlike the interdependence of the delays $\vec{r}$.

3.3. Degenerate case. If $\det \Delta_D(\lambda;\ \vec{r}) \equiv 1$, which occurs, for instance, if all matrices H_k are lower triangular and have zero diagonal, then the zeros of $\det \Delta_N (\lambda;\ \vec{r}, \vec{v})$ are equal to the zeros of

$$Q(\lambda;\ \vec{r}, \vec{v}) := \det \left(\lambda I - \operatorname{adj}(\Delta_D(\lambda;\ \vec{r})) \left(A_0 + \sum_{k=1}^{p_2} A_k e^{-\lambda v_k} \right) \right). \tag{3.3}$$

Equation (3.3) can also be interpreted as the characteristic function of a linear time-delay system of retarded type, of which the spectral properties carry over (see, e.g., [11, 8, 22] for spectral properties of retarded-type systems).

4. Main results, computational expressions for determining strong stability.

The aim of this section is to derive computationally tractable characterizations of the quantities $\bar{C}_D(\vec{r})$ and Ξ, which, by Propositions 3.5 and 3.10, directly result in strong stability conditions. First, we consider special cases where particular conditions are put on the interdependence of the delays. In this way expressions are obtained which directly extend the expressions for the case of independent delays presented in [23], but the derivation is more involved. Next, we show how an *arbitrary* delay dependency of the form (1.2) can be dealt with. The main results will be presented in Theorems 4.3 and 4.7.

4.1. Results for special dependencies in the delays.

We start by stating a technical lemma.

LEMMA 4.1. *Assume that there is a vector* $\vec{\beta} \in (\mathbb{R}_0)^m$ *such that*

$$\vec{\gamma}_k \cdot \vec{\beta} = \vec{\gamma}_l \cdot \vec{\beta} \neq 0 \quad \forall k, l \in \{1, \dots, p_1\}. \tag{4.1}$$

Let $\vec{r} \in (\mathbb{R}_0^+)^m$ *and* $c \in \mathbb{R}$. *If the function*

$$\vec{\theta} \in [0,\ 2\pi]^m \mapsto \alpha \left(-\sum_{k=1}^{p_1} H_k e^{-c\vec{\gamma}_k \cdot \vec{r}} e^{-i\vec{\gamma}_k \cdot \vec{\theta}} \right)$$

has a global maximum, α_0, *for* $\vec{\theta} = \vec{\theta}_0$, *then*

$$\alpha_0 \in \sigma \left(-\sum_{k=1}^{p_1} H_k e^{-c\vec{\gamma}_k \cdot \vec{r}} e^{-i\gamma_k \cdot \vec{\theta}_0} \right).$$

Proof. Let $\lambda(\vec{\theta}_0)$ be an *active* eigenvalue of $\left(-\sum_{k=1}^{p_1} H_k e^{-c\vec{\gamma}_k \cdot \vec{r}} e^{-i\gamma_k \cdot \vec{\theta}_0}\right)$, that is,

$$\Re(\lambda) = \alpha \left(-\sum_{k=1}^{p_1} H_k e^{-c\vec{\gamma}_k \cdot \vec{r}} e^{-i\gamma_k \cdot \vec{\theta}_0} \right).$$

Because the spectral abscissa of a matrix which smoothly depends on parameters is a continuously differentiable function of these parameters in the neighborhood of a global maximum (see [5]), the eigenvalue $\lambda(\vec{\theta}_0)$ is either simple or semisimple. Hence, it defines a continuously differentiable function

$$\vec{\theta} \in \mathcal{B}(\vec{\theta}_0) \mapsto \lambda(\vec{\theta}), \tag{4.2}$$

where $\mathcal{B}(\vec{\theta}_0)$ is some open set of $\mathbb{R}^m$ containing $\vec{\theta}_0$. Let the continuously differentiable functions $\vec{\theta} \mapsto w_0^*(\vec{\theta})$ and $\vec{\theta} \mapsto v_0(\vec{\theta})$ correspond to (normalized) left and right eigenvectors:

$$\left(\lambda(\vec{\theta})I + \sum_{k=1}^{p_1} H_k e^{-c\vec{\gamma}_k\cdot\vec{r}} e^{-i\vec{\gamma}_k\cdot\vec{\theta}}\right) v_0(\vec{\theta}) = 0, \tag{4.3}$$

$$w_0^*(\vec{\theta}) \left(\lambda(\vec{\theta})I + \sum_{k=1}^{p_1} H_k e^{-c\vec{\gamma}_k\cdot\vec{r}} e^{-i\vec{\gamma}_k\cdot\vec{\theta}}\right) = 0, \quad \vec{\theta} \in \mathcal{B}(\vec{\theta}_0). \tag{4.4}$$

Because the spectral abscissa has a maximum at $\vec{\theta}_0$, we have

$$\left.\frac{\partial \Re\lambda(\vec{\theta})}{\partial \theta_j}\right|_{\vec{\theta}=\vec{\theta}_0} = 0, \ j = 1, \ldots, m.$$

Note that

$$\left.\frac{\partial \Re\lambda(\vec{\theta})}{\partial \theta_j}\right|_{\vec{\theta}=\vec{\theta}_0} = \Re \left.\frac{\partial \lambda(\vec{\theta})}{\partial \theta_j}\right|_{\vec{\theta}=\vec{\theta}_0},$$

where $\frac{\partial\lambda(\vec{\theta})}{\partial\theta_j}\big|_{\vec{\theta}=\vec{\theta}_0}$ can be computed by differentiating (4.3) at $\vec{\theta}_0$, premultiplying the result with $w_0^*(\vec{\theta}_0)$ and using (4.4). In this way we arrive at

(4.5)

$$\left.\frac{\partial \Re(\lambda(\vec{\theta}))}{\partial \theta_j}\right|_{\vec{\theta}=\vec{\theta}_0} = \Re \frac{w_0^*(\vec{\theta}_0)\left(\sum_{k=1}^{p_1} \gamma_{k,j} i H_k e^{-c\vec{\gamma}_k\cdot\vec{r}} e^{-i\vec{\gamma}_k\cdot\vec{\theta}_0}\right) v_0(\vec{\theta}_0)}{w_0^*(\vec{\theta}_0) v_0(\vec{\theta}_0)} = 0, \ j = 1, \ldots, m.$$

Let $\vec{\beta} \in (\mathbb{R}_0)^m$ be such that condition (4.1) holds. From (4.5) it follows that

$$\begin{aligned}
0 &= \sum_{j=1}^{m} \beta_j \ \Re\left(\frac{w_0^*\left(\sum_{k=1}^{p_1} \gamma_{k,j} i H_k e^{-c\vec{\gamma}_k\cdot\vec{r}} e^{-i\vec{\gamma}_k\cdot\vec{\theta}_0}\right) v_0}{w_0^* v_0}\right) \\
&= \Re\left(\sum_{j=1}^{m} \beta_j \frac{w_0^*\left(\sum_{k=1}^{p_1} \gamma_{k,j} i H_k e^{-c\vec{\gamma}_k\cdot\vec{r}} e^{-i\vec{\gamma}_k\cdot\vec{\theta}_0}\right) v_0}{w_0^* v_0}\right) \\
&= \Re\left(\frac{w_0^*\left(\sum_{k=1}^{p_1} (\vec{\gamma}_k\cdot\vec{\beta}) \ i H_k e^{-c\vec{\gamma}_k\cdot\vec{r}} e^{-i\vec{\gamma}_k\cdot\vec{\theta}_0}\right) v_0}{w_0^* v_0}\right) \\
&= \Re\left(\frac{w_0^*\left(\sum_{k=1}^{p_1} (\vec{\gamma}_1\cdot\vec{\beta}) \ i H_k e^{-c\vec{\gamma}_k\cdot\vec{r}} e^{-i\vec{\gamma}_k\cdot\vec{\theta}_0}\right) v_0}{w_0^* v_0}\right) \\
&= (\vec{\gamma}_1\cdot\vec{\beta}) \ \Re\left(i\frac{w_0^*\left(\sum_{k=1}^{p_1} H_k e^{-c\vec{\gamma}\cdot\vec{r}} e^{-i\vec{\gamma}_k\cdot\vec{\theta}_0}\right) v_0}{w_0^* v_0}\right) \\
&= (\vec{\gamma}_1\cdot\vec{\beta}) \ \Re\left(i\frac{w_0^*\lambda(\vec{\theta}_0) v_0}{w_0^* v_0}\right) \\
&= -(\vec{\gamma}_1\cdot\vec{\beta}) \ \Im(\lambda(\vec{\theta}_0)).
\end{aligned}$$

We conclude that $\Im(\lambda(\vec{\theta}_0)) = 0$ and $\Re(\lambda(\vec{\theta}_0)) = \alpha_0$. □

The next result states that under condition (4.1), the quantity $\bar{C}_D(\vec{r})$ can be computed from the zeros of a scalar function.

PROPOSITION 4.2. *If* $\det \Delta_D(\lambda;\ \vec{r}) \not\equiv 0$ *and there is a vector* $\vec{\beta} \in (\mathbb{R}_0)^m$ *such that*

$$\vec{\gamma}_k \cdot \vec{\beta} = \vec{\gamma}_l \cdot \vec{\beta} \neq 0 \quad \forall k, l \in \{1, \ldots, p_1\},$$

then for every $\vec{r} \in (\mathbb{R}_0^+)^m$, $\bar{C}_D(\vec{r})$ *is the largest zero of the function*

$$c \in \mathbb{R} \rightarrow f(c;\ \vec{r}) - 1,$$

where

$$f(c;\ \vec{r}) = \max_{\vec{\theta} \in [0,\ 2\pi]^m} \alpha\left(-\sum_{k=1}^{p_1} H_k e^{-c\vec{\gamma}_k \cdot \vec{r}} e^{-i\vec{\gamma}_k \cdot \vec{\theta}}\right). \tag{4.6}$$

Proof. From

$$\begin{aligned} \bar{C}_D(\vec{r}) = \max \Bigg\{ c \in \mathbb{R} :\ \det\left(I + \sum_{k=1}^{p_1} H_k e^{-c\vec{\gamma}_k \cdot \vec{r}} e^{-i\vec{\gamma}_k \cdot \vec{\theta}}\right) = 0 \\ \text{for some } \vec{\theta} \in [0,\ 2\pi]^m \Bigg\} \end{aligned} \tag{4.7}$$

(see Proposition 3.2), it follows that there exists at least one value of c such that $f(c;\ \vec{r}) \geq 1$. As $\lim_{c \rightarrow +\infty} f(c;\ \vec{r}) = 0$, the following number is well defined:

$$\hat{c}(\vec{r}) := \max\{c :\ f(c;\ \vec{r}) = 1\}.$$

It is clear that $f(c;\ \vec{r}) \leq 1$ if $c \geq \hat{c}(\vec{r})$. By (4.7) this implies that

$$\hat{c}(\vec{r}) \geq \bar{C}_D(\vec{r}). \tag{4.8}$$

Next, from Lemma 4.1 and the fact that $f(\hat{c}(\vec{r});\ \vec{r}) = 1$ it follows that there exists a $\vec{\theta}_0(\vec{r}) \in [0,\ 2\pi]^{p_1}$ such that

$$1 \in \sigma\left(-\sum_{k=1}^{p_1} H_k e^{-\hat{c}(\vec{r})\vec{\gamma}_k \cdot \vec{r}} e^{-i\vec{\gamma}_k \cdot \vec{\theta}_0(\vec{r})}\right).$$

By (4.7) one concludes that

$$\bar{C}_D(\vec{r}) \geq \hat{c}(\vec{r}). \tag{4.9}$$

From (4.8) and (4.9) we get $\bar{C}_D(\vec{r}) = \hat{c}(\vec{r})$, which is equivalent to the assertion of the proposition. □

By further imposing that the vector $\vec{\beta}$, appearing in Proposition 4.2, has *positive* components only—among others—an explicit expression for Ξ, and thus an explicit strong stability condition, is obtained.

THEOREM 4.3. *Define*

$$\delta_0 := \max_{\vec{\theta} \in [0,\ 2\pi]^m} \alpha\left(-\sum_{k=1}^{p_1} H_k e^{-i\vec{\gamma}_k \cdot \vec{\theta}}\right). \tag{4.10}$$

If $\det \Delta_D(\lambda;\ \vec{r}) \not\equiv 0$ *and there is a vector* $\vec{\beta} \in (\mathbb{R}_0^+)^m$ *such that*

$$\vec{\gamma}_k \cdot \vec{\beta} = \vec{\gamma}_l \cdot \vec{\beta} \quad \forall k, l \in \{1, \ldots, p_1\},$$

then the assertion of Proposition 4.2, *can be strengthened as follows:*

1. *for all* $\vec{r} \in (\mathbb{R}_0^+)^m$, $\bar{C}_D(\vec{r})$ *is the unique zero of the strictly decreasing function* $c \in \mathbb{R} \mapsto f(c;\ \vec{r}) - 1$, *with* f *given by* (4.6);
2. *we have*

$$\Xi = \operatorname{sign} \log(\delta_0);$$

3. *if* $\delta_0 > 1$, *then there exists a vector* $\vec{r}_0 \in (\mathbb{R}_0^+)^m$ *for which* $\bar{C}_D(\vec{r}_0) > 0$.

Proof. We first prove the second and third statements. According to its definition we evaluate Ξ as

$$\Xi = \operatorname{sign}\left(\bar{C}_D(\vec{\beta})\right). \tag{4.11}$$

From Proposition 4.2 $\bar{C}_D(\vec{\beta})$ is the largest zero of the function

$$c \in \mathbb{R} \mapsto e^{-c\vec{\gamma}_1 \cdot \vec{\beta}} \max_{\vec{\theta} \in [0,\ 2\pi]^m} \alpha\left(-\sum_{k=1}^{p_1} H_k e^{-i\vec{\gamma}_k \cdot \vec{\theta}}\right),$$

thus

$$\bar{C}_D(\vec{\beta}) = \frac{1}{\vec{\gamma}_1 \cdot \vec{\beta}} \log(\delta_0). \tag{4.12}$$

The second and third assertions of the proposition follow from (4.11) and (4.12).

The proof of the first assumption is analogous to the proof of Theorem 6 of [23] and relies on the second assertion, combined with an approximation and continuation argument. ▯

Remark 4.4. If $p_1 = m$ and $\tau_k = r_k,\ 1 \leq k \leq m$, then Proposition 4.3 reduces to Theorem 6 and Proposition 1 of [23] and δ_0 is an equivalent quantity with γ_0 of [13].

4.2. Results for general case: Lifting procedure. Recall that the characteristic function of (2.2) is given by

$$\Delta_D(\lambda;\ \vec{r}) = \det\left(I + \sum_{k=1}^{p_1} H_k e^{-\lambda\ \vec{\gamma}_k \cdot \vec{r}}\right). \tag{4.13}$$

By formally setting

$$x_i = e^{-\lambda\ r_i},\ i = 1, \ldots, m,$$

the function (4.13) can be interpreted as a *multivariable* polynomial

$$p(x_1, \ldots, x_m) := \det\left(I + \sum_{k=1}^{p_1} H_k \left(\Pi_{l=1}^m x_l^{\gamma_{k,l}}\right)\right), \tag{4.14}$$

with some constraints on the variables.

Using results from realization theory, one can show that the polynomial (4.14) can be "lifted" and expressed as the determinant of a (linear) pencil. To do so, we write the polynomial matrix

$$I + \sum_{k=1}^{p_1} H_k \left(\Pi_{l=1}^m x_l^{\gamma_{k,l}}\right)$$

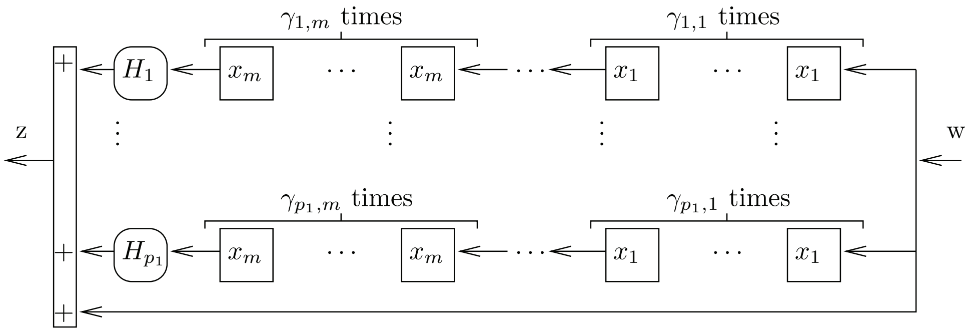


FIG. 4.1. *Block diagram of the relation* (4.15).

as a so-called lower linear fractional representation (see [33]). Let "input" $w \in \mathbb{R}^n$ and "output" $z \in \mathbb{R}^n$ be such that

$$z = \left(I + \sum_{k=1}^{p_1} H_k \left(\Pi_{l=1}^{m} x_l^{\gamma_{k,l}} \right) \right) w. \tag{4.15}$$

This relation can be represented by the block diagram shown in Figure 4.1. By "pulling out" the square blocks, corresponding to the variables, and collecting them in a diagonal matrix, it follows that (4.15) is equivalent to

$$\left[\begin{array}{c} z \\ y \end{array} \right] = M \left[\begin{array}{c} w \\ u \end{array} \right], \quad u = \Delta(x_1, \ldots, x_m)\, y, \tag{4.16}$$

where

$$M = \left[\begin{array}{c||c} M_{11} & M_{12} \\ \hline\hline M_{21} & M_{22} \end{array} \right] := \left[\begin{array}{c||ccc|c|ccc} I & \multicolumn{3}{c|}{\overbrace{0 \cdots 0\ H_1}^{s_1\ \text{blocks}}} & \cdots & \multicolumn{3}{c}{\overbrace{0 \cdots 0\ H_{p_1}}^{s_{p_1}\ \text{blocks}}} \\ \hline\hline I & 0 & \cdots & 0 & & & & \\ 0 & I & & \vdots & & & & \\ \vdots & & \ddots & & & & & \\ 0 & & I & 0 & & & & \\ \hline \vdots & & & & \ddots & & & \\ \hline I & & & & & 0 & \cdots & 0 \\ 0 & & & & & I & & \vdots \\ \vdots & & & & & & \ddots & \\ 0 & & & & & & I & 0 \end{array} \right] \tag{4.17}$$

and

(4.18)

$$\Delta(x_1, \ldots, x_m) = \left[\begin{array}{ccc|c|ccc} x_1 I_{n\gamma_{1,1}} & & & & & & \\ & \ddots & & & & & \\ & & x_m I_{n\gamma_{1,m}} & & & & \\ \hline & & & \ddots & & & \\ \hline & & & & x_1 I_{n\gamma_{p_1,1}} & & \\ & & & & & \ddots & \\ & & & & & & x_m I_{n\gamma_{p_1,m}} \end{array} \right],$$

with $s_k = \sum_{l=1}^{m} \gamma_{k,l}$, $1 \le k \le p_1$, and I_u, $u \in \mathbb{N}$, denoting the u-by-u unity matrix.

From (4.16) we obtain

$$\begin{aligned} z &= \mathcal{F}_l(M, \Delta(x_1, \ldots, x_m))\, y \\ &:= \big(I + M_{12}\Delta(x_1, \ldots, x_m)(I - M_{22}\Delta(x_1, \ldots, x_m))^{-1} M_{21}\big)\, y. \end{aligned}$$

It follows that

$$\begin{aligned} p(x_1, \ldots, x_m) &= \det\big(I + M_{12}\Delta(x_1, \ldots, x_m)(I - M_{22}\Delta(x_1, \ldots, x_m))^{-1} M_{21}\big) \\ &= \det\big(I + (I - M_{22}\Delta(x_1, \ldots, x_m))^{-1} M_{21} M_{12}\Delta(x_1, \ldots, x_m)\big) \\ &= \det\left(I + (M_{21}M_{12} - M_{22})\Delta(x_1, \ldots, x_m)\right) \\ &= \det\Big(I + \textstyle\sum_{k=1}^{m} \tilde{H}_k x_k\Big), \end{aligned}$$

where

$$\tilde{H}_k = (M_{21}M_{12} - M_{22})\Delta(\vec{e}_k), \quad k = 1, \ldots, m, \tag{4.19}$$

and $\vec{e}_k$ is the kth unit vector in $\mathbb{R}^m$. In this way, we arrive at the following result.

Proposition 4.5. *There always exist real square matrices $\tilde{H}_1, \ldots, \tilde{H}_m$ of equal dimensions such that*

$$p(x_1, \ldots, x_m) = \det\left(I + \sum_{k=1}^{m} \tilde{H}_k x_k\right), \tag{4.20}$$

or, equivalently,

$$\det \Delta_D(\lambda;\ \vec{r}) = \det\left(I + \sum_{k=1}^{m} \tilde{H}_k e^{-\lambda r_k}\right). \tag{4.21}$$

A solution is given by (4.19), *where M and Δ are defined in* (4.17) *and* (4.18).

Remark 4.6. The lifting of (4.14) to an expression of the form (4.20) is *not* unique. Furthermore, the presented solution (4.17)–(4.19) does not necessarily correspond to a solution where the matrices $\tilde{H}_k$ have minimal dimensions. In fact, a minimal realization can be obtained from a block diagram representation of (4.15) (possibly different from the one shown in Figure 4.1), where the number of square blocks (thus, the dimension of $\Delta(x_1, \ldots, x_m)$) is minimal. As we shall illustrate with two examples, the construction of such minimal realization strongly depends on the specific properties of the polynomial under consideration and is hard to automate. Notice here that finding an algorithm for the automatic construction of a *minimal* realization is still an *open problem* in realization theory. Note also that the lifting procedure presented above is systematic and generally applicable. For more results on linear fractional representations (LFRs) of multivariable polynomials we refer to the specialized literature; see, e.g., Chapter 10 in [33] for representations coming from state-space realizations in control theory, and Chapter 14 in [15] for many references and extensions to symmetric representations and polynomials with noncommutative variables. See also [20] for an excellent user-friendly publicly available MATLAB toolbox which contains—among other things—routines to compute LFRs and numerical heuristics to reduce the order of LFRs.

We now return to the original problem. From the expression (4.21) it follows that $\Delta_D(\lambda;\ \vec{r})$ can be interpreted as the characteristic function of the "lifted" difference equation

$$\chi(t) + \sum_{k=1}^{m} \tilde{H}_k \chi(t - r_k) = 0.$$

As this equation satisfies the condition assumed in the propositions of section 4.1, the following result directly follows.

THEOREM 4.7. *For the delay difference equation* (2.2) *we have*

$$\Xi = \text{sign}\ \log(\delta_0),$$

where

$$\delta_0 := \max_{\vec{\theta} \in [0,\ 2\pi]^m} \alpha \left(\sum_{k=1}^{m} -\tilde{H}_k e^{-i\theta_k} \right)$$

and the matrices $\tilde{H}_k$ *are such that* (4.21) *in Proposition* 4.5 *holds.*

Furthermore, for all $\vec{r} \in (\mathbb{R}_0^+)^m$, $\bar{C}_D(\vec{r})$ *is the* unique *zero of the* strictly decreasing *function*

$$c \in \mathbb{R} \to f(c;\ \vec{r}) = \max_{\vec{\theta} \in [0,\ 2\pi]^m} \alpha \left(\sum_{k=1}^{m} -\tilde{H}_k e^{-cr_k} e^{-i\theta_k} \right).$$

With two examples we illustrate the lifting procedure for the computation of the matrices $\tilde{H}_k$, $1 \leq k \leq m$, because this is the main step in the application of Theorem 4.7.

Example 4.8. If $p_1 = 3,\ m = 2$, and

$$\vec{\gamma}_1 = (1, 0),\ \vec{\gamma}_2 = (0, 1),\ \vec{\gamma}_3 = (1, 1),$$

then the delay difference equation (2.2) becomes

$$z(t) + H_1 z(t - r_1) + H_2 z(t - r_2) + H_3 z(t - (r_1 + r_2)) = 0. \tag{4.22}$$

This case is not directly covered in section 4.1 since there does not exist a vector $\vec{\beta} \in (\mathbb{R}_0^+)^M$ such that

$$\vec{\gamma}_k \cdot \vec{\beta} = \vec{\gamma}_l \cdot \vec{\beta} \neq 0 \ \ \forall k, l \in \{1, 2, 3\}.$$

The characteristic equation of (4.22) is given by

$$\det \left(I + H_1 e^{-\lambda r_1} + H_2 e^{-\lambda r_2} + H_3 e^{-\lambda(r_1 + r_2)} \right) = 0.$$

An application of Proposition 4.5 leads to the equivalent expression

$$\det \left(I + \begin{bmatrix} H_1 & 0 & 0 & 0 \\ H_1 & 0 & 0 & 0 \\ H_1 & 0 & 0 & 0 \\ 0 & 0 & -I & 0 \end{bmatrix} e^{-\lambda r_1} + \begin{bmatrix} 0 & H_2 & 0 & H_3 \\ 0 & H_2 & 0 & H_3 \\ 0 & H_2 & 0 & H_3 \\ 0 & 0 & 0 & 0 \end{bmatrix} e^{-\lambda r_2} \right) = 0.$$

In Figure 4.2 (top) we show a block diagram of the relation

$$z = (I + H_1 x_1 + H_2 x_2 + H_3 x_1 x_2)\, w, \tag{4.23}$$

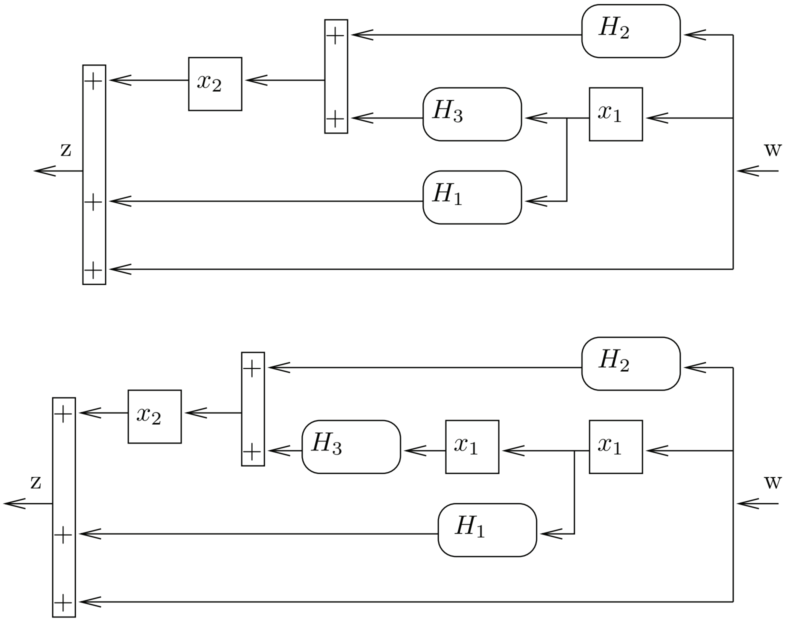


FIG. 4.2. *Block diagram representation of* (4.23) *(above) and* (4.25) *(below), using a minimum number of square blocks.*

where we have minimized the number of square blocks (corresponding to a variable), that is, we have minimized the dimension of $\Delta(x_1, x_2)$. It leads to the minimal order[3] lifting, given by

$$\det\left(I + \underbrace{\begin{bmatrix} H_1 & 0 \\ H_2H_1 - H_3 & 0 \end{bmatrix}}_{\tilde{H}_1} e^{-\lambda r_1} + \underbrace{\begin{bmatrix} 0 & I \\ 0 & H_2 \end{bmatrix}}_{\tilde{H}_2} e^{-\lambda r_2}\right) = 0. \tag{4.24}$$

Example 4.9. If $p_1 = 3,\ m = 2$, and

$$\vec{\gamma}_1 = (1,0),\ \vec{\gamma}_2 = (0,1),\ \vec{\gamma}_3 = (2,1),$$

then the characteristic equation of (2.2) becomes

$$\det\left(I + H_1 e^{-\lambda r_1} + H_2 e^{-\lambda r_2} + H_3 e^{-\lambda(2r_1+r_2)}\right) = 0.$$

The systematic lifting procedure proposed in Proposition 4.7 leads us to the equivalent expression

$$\det\left(I + \begin{bmatrix} H_1 & 0 & 0 & 0 & 0 \\ H_1 & 0 & 0 & 0 & 0 \\ H_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -I & 0 & 0 \\ 0 & 0 & 0 & -I & 0 \end{bmatrix} e^{-\lambda r_1} + \begin{bmatrix} 0 & H_2 & 0 & 0 & H_3 \\ 0 & H_2 & 0 & 0 & H_3 \\ 0 & H_2 & 0 & 0 & H_3 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} e^{-\lambda r_2}\right) = 0.$$

[3] without assumptions on the matrices H_k. A further reduction may be possible when the matrices H_k are specified or information about their structure is present.

A minimal order lifting follows from the block diagram representation of

$$z = (I + H_1 x_1 + H_2 x_2 + H_3 x_1 x_2)\, w, \tag{4.25}$$

shown in Figure 4.2 (bottom), and it is given by

$$\det\left(I + \underbrace{\begin{bmatrix} H_1 & 0 & 0 \\ -I & 0 & 0 \\ H_2 H_1 & -H_3 & 0 \end{bmatrix}}_{\tilde{H}_1} e^{-\lambda r_1} + \underbrace{\begin{bmatrix} 0 & 0 & I \\ 0 & 0 & 0 \\ 0 & 0 & H_2 \end{bmatrix}}_{\tilde{H}_2} e^{-\lambda r_2} \right) = 0. \tag{4.26}$$

Finally, we illustrate that the lifting step is necessary if the assumption on the interdependency of the delays of Proposition 4.3 is not satisfied.

Example 4.10. When applying Theorem 4.7 to the delay difference equation

$$z(t) + \frac{20}{101} z(t - r_1) - \frac{40}{101} z(t - r_2) - \frac{80}{101} z(t - (r_1 + r_2)) = 0,$$

for which the lifting (4.24) can be used, we get $\delta_0 = 0.9945 < 1$, thus $\Xi < 0$, and we can conclude strong stability. On the other hand, formula (4.10) would result in $\delta_0 = 1.0066 > 1$. This demonstrates that lifting may be necessary if the assumption of Proposition 4.3 is not satisfied, and that the assertions of Proposition 4.3 are not condensed formulations of the assertions of Theorem 4.7.

5. Illustrations and applications.

5.1. Numerical example. We apply the theoretical results derived above to the system

$$\frac{d}{dt}\left(x(t) + \sum_{k=1}^{3} H_k x(t - \tau_k) \right) = A_0 x(t) + A_1 x(t - \nu_1), \tag{5.1}$$

where the system matrices are given by

$$\begin{gathered} H_1 = \begin{bmatrix} \frac{1}{2} & 0 \\ -\frac{1}{8} & \frac{1}{2} \end{bmatrix}, \quad H_2 = \begin{bmatrix} -\frac{1}{8} & 1 \\ -\frac{1}{4} & \frac{1}{4} \end{bmatrix}, \quad H_3 = \begin{bmatrix} -\frac{1}{8} & -\frac{1}{4} \\ 0 & \frac{1}{8} \end{bmatrix}, \\ A_0 = \begin{bmatrix} 0 & -\frac{11}{40} \\ \frac{11}{80} & 0 \end{bmatrix}, \quad A_1 = \begin{bmatrix} -\frac{1}{64} & -\frac{1}{8} \\ -\frac{1}{8} & -\frac{1}{32} \end{bmatrix} \end{gathered} \tag{5.2}$$

and the dependency of the delays is described by

$$\tau_1 = r_1, \ \ \tau_2 = r_2, \ \ \tau_3 = 2r_1 + r_2, \tag{5.3}$$

with r_1 and r_2 independent.

In Figure 5.1 we show the rightmost characteristic roots of (5.1)–(5.3) for $(r_1, r_2) = (1, 2)$ and $\nu_1 = 1$, computed with the quasi-polynomial mapping–based rootfinder (QPMR) [30]. Note that the exponentially transformed characteristic roots correspond to the eigenvalues of the operator $\mathcal{T}(1;\ (r_1, r_2), \upsilon_1)$. We have

$$c_N((1, 2), 1) = -0.025 \ \text{ and } \ c_D((1, 2)) = -0.191.$$

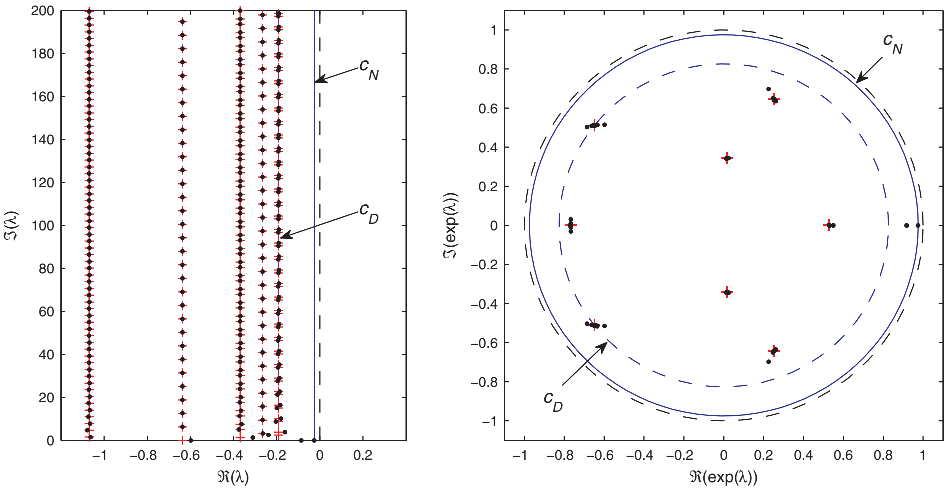


FIG. 5.1. *Rightmost characteristic roots of the system* (5.1)–(5.3) *with* $(r_1, r_2) = (1, 2)$ *and* $\nu_1 = 1$. *Dots—characteristic roots of the neutral system; crosses—characteristic roots of the associated difference equation.*

Let us remark that the latter quantity can be calculated from the zeros of a polynomial, because

$$\Delta_D(\lambda;\ (1,2)) = \det(I + H_1\chi + H_2\chi^2 + H_3\chi^4),$$

provided $\chi = e^{-\lambda}$. Thus, if the characteristic roots of the delay difference equation with the commensurate delays are exponentially transformed, they are mapped to a finite number of points. Due to the relation

$$\sigma_e(\mathcal{T}(t;\ (r_1, r_2), \nu_1)) = \sigma(\mathcal{T}_D(t;\ (r_1, r_2))),$$

the transformed roots of the neutral system accumulate to these points. This can be seen in the right frame of Figure 5.1.

In order to show the effect of small delay perturbations, we depict in Figure 5.2 the characteristic roots of (5.1)–(5.3) for $(r_1, r_2) = (1, 2 + \pi/100)$ and $\nu_1 = 1$. We also indicate the quantity

$$\bar{C}_D((1,2), \nu_1) = -0.066,$$

which can be computed by applying Theorem 4.7, starting from the representation (4.26). The fact that $\bar{C}_D((1,2)) > c_D((1,2))$ illustrates the noncontinuity of the function $\vec{r} \mapsto c_D(\vec{r})$. Notice from Figures 5.1–5.2 that in any right half plane $\{\lambda \in \mathbb{C} : \Re(\lambda) > \bar{C}_D + \epsilon\}$, $\epsilon > 0$, the neutral equation has only a finite number of characteristic roots.

Because $\bar{C}_D((1,2)) < 0$, which implies $\Xi < 0$, and $c_D((1,2),1) < 0$, the null solution of (5.1)–(5.3) is strongly exponentially stable.

If one is only interested in checking strong stability of the delay difference equation, then according to Theorem 4.7 it is sufficient to check whether $\delta_0 < 1$, where

$$\delta_0 = \max_{\vec{\theta} \in [0,\ 2\pi]^2} \alpha\left(-\tilde{H}_1 e^{-i\theta_1} - \tilde{H}_2 e^{-i\theta_2}\right), \tag{5.4}$$

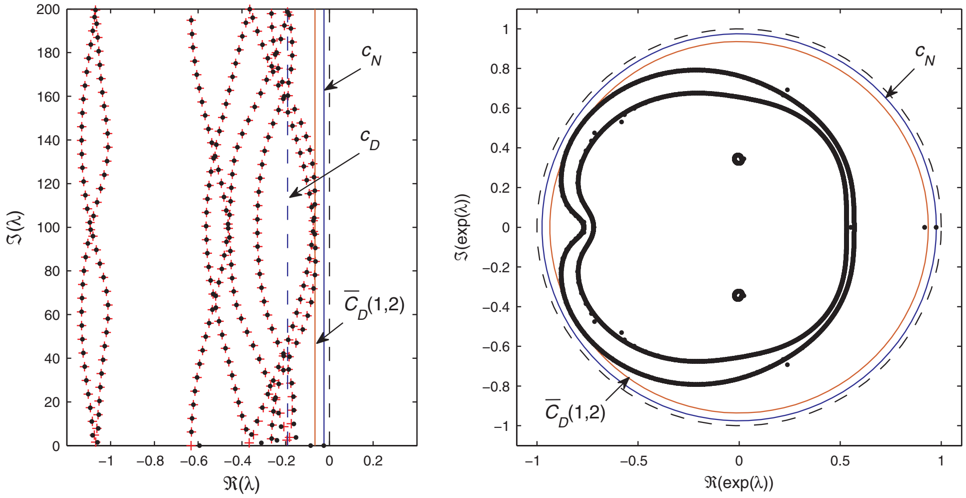


FIG. 5.2. *Rightmost characteristic roots of the system* (5.1)–(5.3) *with* $(r_1, r_2) = (1, 2 + \frac{\pi}{100})$ *and* $\nu_1 = 1$.

with $\tilde{H}_1, \tilde{H}_2$ defined in (4.26). From (5.4) we get

$$\delta_0 = \max_{\theta \in [0,2\pi]} \alpha(-\tilde{H}_1 - \tilde{H}_2 e^{-i\theta}) = 0.901.$$

In Figure 5.3 we show contour lines of the spectral abscissa function

$$(\theta_1, \theta_2) \mapsto \alpha(-\tilde{H}_1 e^{-i\theta_1} - \tilde{H}_2 e^{-i\theta_2}), \tag{5.5}$$

as well as curves corresponding to the values of θ_1 and θ_2 for which a rightmost eigenvalue of

$$-\tilde{H}_1 e^{-i\theta_1} - \tilde{H}_2 e^{-i\theta_2} \tag{5.6}$$

is real. As can be seen from the figure, the matrix (5.6) has a real rightmost eigenvalue if (θ_1, θ_2) is a global maximizer of (5.5). This is in accordance with the statement of Lemma 4.1.

Finally, let us illustrate that the effect of delay perturbations strongly depends on the interdependence of the delays. If, instead of the relation (5.3), we assume that the delays τ_k, $1 \le k \le 3$, in (5.1) can vary independently of each other independent, that is,

$$\tau_k = r_k, \ k = 1, \dots, 3,$$

then we get

$$\bar{C}_D((1, 2, 4)) = 0.055,$$

which shows that strong stability is lost. Note for comparison that with the previously considered dependency structure (5.3) the nominal values $\vec{r} = (1, 2)$ also corresponded to $\vec{\tau} = (1, 2, 4)$.

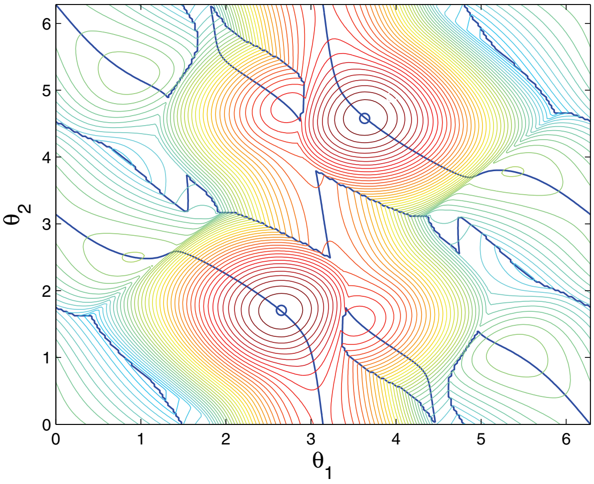


FIG. 5.3. *Contour lines of the function* (5.5). *The global maxima are indicated with "○". The dark curves correspond to values of θ_1 and θ_2 for which the rightmost eigenvalue of* (5.6) *is real.*

5.2. Boundary-controlled partial differential equation. The following model from [19] (see also [6, 24] for a simplified version) describes movement of a string fixed at one side and controlled by changing the direction of the external force at the other side:

$$w_{tt}(x,t) - w_{xx}(x,t) + 2aw_t(x,t) + a^2 w(x,t) = 0,\ t \geq 0, x \in [0,\ 1], \tag{5.7}$$

$$w(0,t) = 0, \quad w_x(1,t) = -kw_t(1,t-h). \tag{5.8}$$

The variable $w(x,t)$ describes the movement at position x at time t. The parameter $h \geq 0$ represents a small delay in the velocity feedback, $k \geq 0$ is the controller gain, and $a \geq 0$ represents a damping constant.

When substituting a solution of the form $w(x,t) = e^{\lambda t} z(x)$ in (5.7)–(5.8) the following characteristic equation is obtained:

$$1 + e^{-2a} e^{-\lambda 2} + ke^{-\lambda h} - ke^{-2a} e^{-\lambda(2+h)} = 0. \tag{5.9}$$

Note that this equation can be interpreted as the characteristic equation of a delay difference equation of the form (2.2), exhibiting three delays $(\tau_1, \tau_2, \tau_3) = (2, h, 2+h)$ that depend on two independent delays $(r_1, r_2) = (2, h)$.

If $h = 0$, the characteristic roots are

$$\lambda = -\frac{1}{2} \log \left| \frac{1+k}{1-k} \right| - a + i \left(\pi l + \frac{\pi}{4}(1 + \operatorname{sign}(k-1)) \right), \quad l \in \mathbb{Z}.$$

As for all $k \neq 1$,

$$c(k) := -\frac{1}{2} \log \left| \frac{1+k}{1-k} \right| - a < 0, \tag{5.10}$$

the system with $h = 0$ is stable for all $k \neq 1$. As k approaches 1, the real parts of the characteristic roots move off to $-\infty$, which indicates superstability at $k = 1$ (meaning that perturbations disappear in a finite time). This is indeed the case and can be explained as follows: the general solution of (5.7) can be written as a combination of two traveling waves: a solution $\phi(x-t)e^{-at}$ moving to the right and a solution $\psi(x+t)e^{-at}$ moving to the left. If $k=1$, then $\phi(x-t)e^{-at}$ satisfies the second boundary condition, and thus the reflection coefficient at $x=1$ is zero; at $x=0$ the wave $\phi(x+t)$ is reflected completely. Consequently all perturbations of the zero solution disappear in a finite time (at most 2 time units).

Next, we look at the effect of a small feedback delay h in the application of the boundary control. If the delays $(r_1, r_2) = (2, h)$ are rationally independent, which occurs if h is an irrational number, then we have $c_D(\vec{r}) = \bar{C}_D(\vec{r})$ (Proposition 3.2), and the stability condition is given by $\Xi < 0$ (which also guarantees stability for all $h > 0$). To compute Ξ, we apply Theorem 4.7, based on the lifting (4.24). This yields

$$\begin{aligned}
\delta_0 &= \max\nolimits_{(\theta_1,\theta_2)\in[0,\ 2\pi]^2}\ \alpha\left(-\begin{bmatrix} e^{-2a} & 0 \\ 2ke^{-2a} & 0 \end{bmatrix} e^{-i\theta_1} - \begin{bmatrix} 0 & 1 \\ 0 & k \end{bmatrix} e^{-i\theta_2}\right) \\
&= \max\nolimits_{\theta\in[0,\ 2\pi]}\ r_\sigma\left(\begin{bmatrix} e^{-2a} & 0 \\ 2ke^{-2a} & 0 \end{bmatrix} + \begin{bmatrix} 0 & 1 \\ 0 & k \end{bmatrix} e^{-i\theta}\right) \\
&= \max\left\{|\lambda| :\ 1 - \frac{k(\lambda + e^{-2a})}{\lambda^2 - e^{-2a}\lambda} e^{i\theta} = 0,\ \theta \in [0,\ 2\pi], \lambda \in \mathbb{C}\right\} \\
&= \max\left\{|\lambda| :\ \left|\frac{k(\lambda + e^{-2a})}{\lambda^2 - e^{-2a}\lambda}\right| = 1,\ \lambda \in \mathbb{C}\right\} \\
&= \max\left\{|\lambda| :\ \frac{k\left(1 + \frac{e^{-2a}}{|\lambda|}\right)}{|\lambda - e^{-2a}|} = 1,\ \lambda \in \mathbb{C}\right\} \\
&= \frac{1}{2}\left(e^{-2a} + k + \sqrt{(e^{-2a} + k)^2 + 4ke^{-2a}}\right).
\end{aligned}$$

It follows that

$$\Xi = \operatorname{sign}\log(\delta_0) < 0 \Leftrightarrow k < \tanh(a),$$

where $<$ can be replaced with $>, =$. We conclude with the following:

1. If $k < \tanh(a)$, then the system (5.7)–(5.8) is exponentially stable for all $h \geq 0$.
2. If $k > \tanh(a)$, then the system (5.7)–(5.8) is exponentially unstable for all irrational values of h. Consequently, there exist *arbitrarily small* values of h that destroy the exponential stability of the system without delay in the boundary control.

5.3. Delay robustness of state derivative feedback control. In [1, 2] the problem of stabilization and control of the linear system

$$\dot{x}(t) = Ax(t) + Bu(t), \tag{5.11}$$

where $x(t) \in \mathbb{R}^n$ is the vector of state variables, $u \in \mathbb{R}^{n_u}(t)$ is the vector of inputs, and A, B are constant coefficient matrices of compatible dimension, has been solved by the state derivative feedback controller

$$u(t) = -K_d\dot{x}(t). \tag{5.12}$$

The use of state derivative control law is motivated by its easy implementation in applications where accelerometers are used for measuring the system motion, e.g., applications in vibration control, where the state variables typically correspond to positions and velocities. In [1, 2], it is shown that if the system (5.11) is controllable, and $\det(A) \neq 0$, then all the characteristic roots of the closed-loop system can be assigned at arbitrary positions in $\mathbb{C} \setminus \{0\}$. However, results described in [31] indicate that stability of the state derivative feedback control may not be robust against small feedback delays. This issue is investigated in what follows.

If we assume that there is a delay τ_{u_k} on the kth component of input u, $1 \leq k \leq n_u$, and a delay τ_{x_l} in the measurement of the lth component of $\dot{x}$, $1 \leq l \leq n$, then the closed-loop system (5.11)–(5.12) becomes

$$\dot{x}(t) + \sum_{k=1}^{n_u} BE_k \sum_{l=1}^{n} K_d F_l \dot{x}(t - \tau_{u_k} - \tau_{x_l}) = Ax(t), \tag{5.13}$$

where $E_k = [e^k_{i,j}] \in \mathbb{R}^{n_u \times n_u}$ and $F_l = [f^l_{i,j}] \in \mathbb{R}^{n \times n}$ satisfy

$$e^k_{i,j} = \begin{cases} 1, & i = j = k, \\ 0, & \text{otherwise}, \end{cases} \qquad f^l_{i,j} = \begin{cases} 1, & i = j = l, \\ 0, & \text{otherwise} \end{cases}$$

for $k = 1, \ldots, n_u$ and $l = 1, \ldots, n$. Equation (5.13) is of the general form (1.1), provided that we set

$$\begin{aligned} & p_1 = n_u n, \quad p_2 = 0, \quad m = n_u + n, \\ & (\tau_1, \ldots, \tau_{p_1}) = (\tau_{u_1} + \tau_{x_1}, \ldots, \tau_{u_1} + \tau_{x_n}, \ldots, \tau_{u_{n_u}} + \tau_{x_1}, \tau_{u_{n_u}} + \tau_{x_n}), \\ & (r_1, \ldots, r_m) = (\tau_{u_1}, \ldots, \tau_{u_{n_u}}, \tau_{x_1}, \ldots, \tau_{x_n}), \end{aligned} \tag{5.14}$$

and we define vectors $\vec{\gamma}_k$, $1 \leq k \leq p_1$, and matrices A_0, H_k, $1 \leq k \leq p_1$, accordingly. We have the following result.

Proposition 5.1. *Assume the system* (5.11) *is stabilized with the control law* (5.12).

If the feedback gain K_d is such that

$$\gamma_0(K_d) := \max \left\{ \alpha \left(-\sum_{k=1}^{n_u} BE_k \sum_{l=1}^{n} K_d F_l e^{i(\mu_k + \nu_l)} \right) : \vec{\mu} \in [0,\ 2\pi]^{n_u},\ \vec{\nu} \in [0,\ 2\pi]^n \right\} < 1,$$

then the exponential stability of the closed-loop system is robust against small feedback delays.

If $\gamma_0(K_d) \geq 1$, then the exponential stability of the closed-loop stability is not robust against small delay perturbations.

Proof. The interdependence between the delays of the neutral system (5.13) satisfies the condition of Proposition 4.3. Furthermore, for this system the quantity δ_0, defined in Proposition 4.3, reduces to $\gamma_0(K_d)$. Consequently, if $\gamma_0(K_d) < 1$, then $\Xi < 0$. By the bounds on the characteristic roots given in Lemma 3.7, the continuity of the individual characteristic roots w.r.t. the delay parameters and the exponential stability of the delay-free system, we conclude that $c_D(\vec{r},\ \vec{v}) < 0$ for sufficiently small values of $\vec{r}$ and $\vec{v}$. Robustness of stability follows. If $\gamma_0(K_d) > 1$, then the null solution of (5.13) is not strongly exponentially stable, which implies that infinitesimal perturbations on the (arbitrarily small) delays destroy exponential stability. □

6. Conclusions. The stability theory for neutral equations and delay difference equation subjected to delay perturbations has been developed for the case where the delays have an arbitrary dependency structure, with the emphasis on spectral properties and computational expressions for $\bar{C}_D$ and Ξ that, among others, lead to explicit strong stability conditions.

Instrumental to this, it has been shown that a general delay difference equation with dependent delays can always be transformed, without changing the characteristic equation, into a delay difference equation with possibly larger dimension but with independent delays, such that the stability theory for systems with independent delays can be applied to complete the theory. An essential step of the constructive procedure consists of representing a multivariate polynomial as the determinant of a pencil. In this sense it is remarkable how the realization theory, commonly used in robust control and optimization, has proven its usefulness to the problems considered in the paper, which are of a different nature. In addition special cases have been addressed for which the lifting step, which may increase the computational complexity, can be omitted.

More specifically the main results are presented in Theorem 4.3, holding for a special dependency of the delays, and Theorem 4.7, holding for the general case. Theorem 4.7 depends on a lifting of the characteristic function for which Proposition 4.5 guarantees the existence and provides a *constructive* solution.

The results derived in the paper have been applied to various problems, including the study of the effects of unmodeled delays on the stability of a boundary-controlled hyperbolic partial differential equation and of a control scheme involving state derivative feedback, being of importance in vibration control applications. These examples illustrate the importance of taking into account small delays or delay perturbations, as well as the dependency structure of the delays.

Appendix A. Proof of Lemma 3.7. Because $\Delta_D(\lambda;\ \vec{r})$ is invertible, we can write the characteristic equation in the form

$$\det\left(\lambda I - \Delta_D(\lambda;\ \vec{r})^{-1}\left(A_0 + \sum_{k=1}^{p_2} A_k e^{-\lambda v_k}\right)\right) = 0.$$

This equation can be interpreted as

$$\lambda \in \sigma\left(\Delta_D(\lambda;\ \vec{r})^{-1}\left(A_0 + \sum_{k=1}^{p_2} A_k e^{-\lambda v_k}\right)\right),$$

which implies

$$|\lambda| \leq \left\|\Delta_D(\lambda;\ \vec{r})^{-1}\left(A_0 + \sum_{k=1}^{p_2} A_k e^{-\lambda v_k}\right)\right\|.$$

By further working out the estimate, we arrive at the assertion. □

Appendix B. Proof of Proposition 3.8. We prove continuity at $(\vec{r},\vec{v}) = (\vec{r}_0,\vec{v}_0)$, where we consider two cases.

Case 1. $\bar{C}_D(\vec{r}_0) \geq c_N(\vec{r}_0,\vec{v}_0)$.

The proof is by contradiction. By item (1) of Proposition 3.2 a violation of the continuity property would imply the existence of sequences $\left\{\vec{r}^{(\varrho)}\right\}_{\varrho\geq 1}$, $\left\{\vec{v}^{(\varrho)}\right\}_{\varrho\geq 1}$ and the existence of a number $\epsilon > 0$ such that

$$\lim_{\varrho\to\infty} \vec{r}^{(\varrho)} = \vec{r}_0,\ \lim_{\varrho\to\infty} \vec{v}^{(\varrho)} = \vec{v}_0$$

and

$$c_N(\vec{r}^{(\varrho)}, \vec{v}^{(\varrho)}) \geq \bar{C}_D(\vec{r}_0) + \epsilon \ \ \forall \varrho \geq 1.$$

As a consequence, there exists a sequence of complex numbers $\{\lambda^{(\varrho)}\}_{\varrho \geq 1}$ satisfying

$$\Delta_N(\lambda^{(\varrho)};\ \vec{r}^{(\varrho)}, \vec{v}^{(\varrho)}) = 0,\ \Re(\lambda^{(\varrho)}) > \bar{C}_D(\vec{r}_0) + \epsilon/2 \ \ \forall \varrho \geq 1.$$

By Lemma 3.7, there is a compact subset of $\mathbb{C}$ which contains all elements of the sequence $\{\lambda^{(\varrho)}\}_{\varrho \geq 1}$. Consequently, this sequence has at least one accumulation point $\hat{\lambda}$. From Rouché's theorem it follows that

$$\Delta_N(\hat{\lambda};\ \vec{r}_0, \vec{v}_0) = 0.$$

Because $\Re(\hat{\lambda}) > \bar{C}_D(\vec{r}_0)$, we arrive at $c_N(\vec{r}_0, \vec{v}_0) > \bar{C}_D(\vec{r}_0)$ and have a contradiction.

Case 2. $\bar{C}_D(\vec{r}_0) < c_N(\vec{r}_0, \vec{v}_0)$.

Let $\epsilon > 0$ be such that $\bar{C}_D(\vec{r}_0)+\epsilon < c_N(\vec{r}_0, \vec{v}_0)$ and $\Delta_N(\bar{C}_D(\vec{r}_0)+\epsilon+j\omega;\ \vec{r}_0, \vec{v}_0) \neq 0$ for all $\omega \geq 0$. From Lemma 3.7 one concludes that the number of zeros of Δ_N in the right half plane

$$\mathcal{H} := \{\lambda \in \mathbb{C} :\ \Re(\lambda) > \bar{C}_D(\vec{r}_0) + \epsilon\}$$

is finite and invariant for $\|\vec{r} - \vec{r}_0\| < \delta$ and $\|v - v_0\| < \delta$, with δ sufficiently small. The assertion is a consequence of the continuity of the zeros of Δ_N in the half place $\mathcal{H}$ w.r.t. the delay parameters $\vec{r}, \vec{v}$ and the continuity of $\bar{C}_D$ w.r.t. $\vec{r}$. □

REFERENCES

[1] T. H. S. Abdelaziz and M. Valášek, *Direct algorithm for pole placement by state-derivative feedback for multi-input linear systems—nonsingular case*, Kybernetika, 41 (2005), pp. 637–660.

[2] T. H. S. Abdelaziz and M. Valášek, *State derivative feedback by LQR for linear time-invariant systems*, in Proceedings of the 16th IFAC World Congress, Prague, Czech Republic, 2005.

[3] C. E. Avellar and J. K. Hale, *On the zeros of exponential polynomials*, J. Math. Anal. Appl., 73 (1980), pp. 434–452.

[4] A. Bellen, N. Guglielmi, and A. E. Ruehli, *Methods for linear systems of circuit delay differential equations of neutral type*, IEEE Trans. Circuits Syst. I, 76 (1999), pp. 212–215.

[5] J. Burke and M. Overton, *Differential properties of the spectral abscissa and the spectral radius for analytic matrix valued mappings*, Nonlinear Anal., 23 (1994), pp. 467–488.

[6] R. Datko, *Two examples of ill-posedness with respect to time delays revisited*, IEEE Trans. Automat. Control, 42 (1997), pp. 434–452.

[7] K. Engelborghs, M. Dambrine, and D. Roose, *Limitations of a class of stabilization methods for delay equations*, IEEE Trans. Automat. Control, 46 (2001), pp. 336–339.

[8] K. Gu, V. L. Kharitonov, and J. Chen, *Stability of Time-Delay Systems*, Birkhäuser, Boston, 2003.

[9] J. K. Hale, *Parametric stability in difference equations*, Boll. Un. Mat. Ital., 11 (1975), pp. 209–214.

[10] J. K. Hale, *Effects of delays on dynamics*, in Topological Methods in Differential Equations and Inclusions, A. Granas, M. Frigon, and G. Sabidussi, eds., Kluwer Academic, Dordrecht, The Netherlands, 1995, pp. 191–238.

[11] J. K. Hale and S. M. Verduyn Lunel, *Introduction to Functional Differential Equations*, Appl. Math. Sci. 99, Springer-Verlag, New York, 1993.

[12] J. K. Hale and S. M. Verduyn Lunel, *Effects of small delays on stability and control*, Oper. Theory Adv. Appl., 122 (2001), pp. 275–301.

[13] J. K. Hale and S. M. Verduyn Lunel, *Strong stabilization of neutral functional differential equations*, IMA J. Math. Control Inform., 19 (2002), pp. 5–23.

[14] J. K. Hale and S. M. Verduyn Lunel, *Stability and control of feedback systems with time delays*, Internat. J. Systems Sci., 34 (2003), pp. 497–504.

[15] J. Helton, S. A. McCullough, and V. Vinnikov, *Noncommutative convexity arises from linear matrix inequalities*, J. Funct. Anal., 240 (2006), pp. 105–191.

[16] D. Henry, *Linear autonomous neutral functional differential equations*, J. Differential Equations, 19 (1974), pp. 106–128.

[17] V. B. Kolmanovskii and A. Myshkis, *Introduction to the Theory and Applications of Functional Differential Equations*, Math. Appl. 463, Kluwer Academic, Dordrecht, The Netherlands, 1999.

[18] H. Logemann and S. Townley, *The effect of small delays in the feedback loop on the stability of neutral systems*, Systems Control Lett., 27 (1996), pp. 267–274.

[19] H. Logemann, R. Rebarber, and G. Weiss, *Conditions for robustness and nonrobustness of the stability of feedback control systems with respect to small delays in the feedback loop*, SIAM J. Control Optim., 34 (1996), pp. 572–600.

[20] J.-F. Magni, *User Manual of the Linear Fractional Representation Toolbox*, Version 2.0, Technical Report TR 5/10403.01F DCSC, ONERA, Systems, Control and Flight Dynamics Department, Toulouse, France, 2005.

[21] W. Melvin, *Stability properties of functional difference equations*, J. Math. Anal. Appl., 48 (1974), pp. 749–763.

[22] W. Michiels and S.-I. Niculescu, *Stability and Stabilization of Time-Delay Systems. An Eigenvalue-Based Approach*, Adv. Des. Control 12, SIAM, Philadelphia, 2007.

[23] W. Michiels and T. Vyhlídal, *An eigenvalue based approach for the stabilization of linear time-delay systems of neutral type*, Automatica, 41 (2005), pp. 991–998.

[24] W. Michiels, K. Engelborghs, D. Roose, and D. Dochain, *Sensitivity to infinitesimal delays in neutral equations*, SIAM J. Control Optim., 40 (2001), pp. 1134–1158.

[25] S. Mondié and W. Michiels, *Finite spectrum assignment of unstable time-delay systems with a safe implementation*, IEEE Trans. Automat. Control, 48 (2003), pp. 2207–2212.

[26] R. M. Murray, C. A. Jacobson, R. Casas, A. I. Khibnik, C. R. Johnson Jr., B. R., A. A. Peracchio, and W. M. Proscia, *System identification for limit cycling systems: A case study for combustion instabilities*, in Proceedings of the 1998 American Control Conference, Philadelphia, 1998, pp. 2004–2008.

[27] S.-I. Niculescu and B. Brogliato, *Force measurements time-delays and contact instability phenomenon*, European J. Control, 5 (1999), pp. 279–289.

[28] N. Olgac, T. Vyhlídal, and R. Sipahi, *A new perspective in the stability assessment of neutral systems with multiple and cross-talking delays*, SIAM J. Control Optim., 47 (2008), pp. 327–344.

[29] L. Pandolfi, *Stabilization of neutral functional differential equations*, J. Optim. Theory Appl., 20 (1976), pp. 191–204.

[30] T. Vyhlídal and P. Zítek, *Quasipolynomial mapping based rootfinder for analysis of time delay systems*, in Proceedings of the Fourth IFAC Workshop on Time-Delay Systems, Rocquencourt, France, 2003.

[31] T. Vyhlídal, P. McGahan, W. Michiels, and P. Zítek, *Stability impact of the latency in the state derivative feedback*, in Proceedings of the 16th International Conference on Process Control, Strbske Pleso, Slovakia, 2007.

[32] T. Vyhlídal, W. Michiels, P. Zítek, and P. McGahan, *Stability impact of small delays in proportional-derivative state feedback*, Control Engineering Practice, to appear.

[33] K. Zhou, J. C. Doyle, and K. Glover, *Robust and Optimal Control*, Prentice–Hall, Englewood Cliffs, NJ, 1995.

SIAM J. Control Optim.
Vol. 48, No. 2, pp. 787–811

© 2009 Society for Industrial and Applied Mathematics

GENERAL PROJECTIVE SPLITTING METHODS FOR SUMS OF MAXIMAL MONOTONE OPERATORS*

JONATHAN ECKSTEIN[†] AND B. F. SVAITER[‡]

Abstract. We describe a general projective framework for finding a zero of the sum of n maximal monotone operators over a real Hilbert space. Unlike prior methods for this problem, we neither assume $n = 2$ nor first reduce the problem to the case $n = 2$. Our analysis defines a closed convex *extended solution set* for which we can construct a separating hyperplane by individually evaluating the resolvent of each operator. At the cost of a single, computationally simple projection step, this framework gives rise to a family of splitting methods of unprecedented flexibility: numerous parameters, including the proximal stepsize, may vary by iteration and by operator. The order of operator evaluation may vary by iteration and may be either serial or parallel. The analysis essentially generalizes our prior results for the case $n = 2$. We also include a relative error criterion for approximately evaluating resolvents, which was not present in our earlier work.

Key words. monotone operators, splitting algorithms, proximal algorithms

AMS subject classifications. Primary, 49M27; Secondary, 47H05, 47J25, 65J20

DOI. 10.1137/070698816

1. Background and introduction.

This paper considers the inclusion

$$0 \in T_1(x) + \cdots + T_n(x), \tag{1.1}$$

where $n \geq 2$ and $T_1, \ldots, T_n$ are set-valued maximal monotone operators on some real Hilbert space $\mathcal{H}$. Our interest is in *splitting* methods for this problem: iterative algorithms which may evaluate the individual operators T_i or (perhaps approximately) their resolvents $(I + \lambda T_i)^{-1}$, $\lambda > 0$, at various points in $\mathcal{H}$ but never resolvents of sums of the T_i. The idea is that (1.1) has been formulated so that each individual T_i has some relatively convenient structure, but sums of two or more of the T_i might not. Thus, we seek iterative decomposition algorithms that evaluate only the "easy" resolvents $(I + \lambda T_i)^{-1}$ and not "difficult" compound resolvents such as $(I + \lambda(T_i + T_j))^{-1}$, $i \neq j$.

Algorithms of this form have been studied since the 1970s [16], although their roots in numerical methods for single-valued and in particular linear mappings are much older [8, 22]. In the extensive literature of these methods, the case $n = 2$ predominates. The most attractive convergence theory among $n = 2$ algorithms belongs to the Peaceman–Rachford and Douglas–Rachford class, which form a single related family using iterations of the form

$$s^{k+1} = \left(1 - \frac{\rho_k}{2}\right) s^k + \frac{\rho_k}{2} R_{\lambda T_1}\left(R_{\lambda T_2}\left(s^k\right)\right), \tag{1.2}$$

*Received by the editors August 1, 2007; accepted for publication (in revised form) August 18, 2008; published electronically February 25, 2009.
http://www.siam.org/journals/sicon/48-2/69881.html

[†] Department of Management Science and Information Systems and RUTCOR, Rutgers University, 640 Bartholomew Road, Busch Campus, Piscataway, NJ 08854 (jeckstei@rci.rutgers.edu). This author's research was supported in part by a faculty research grant from Rutgers Business School—Newark and New Brunswick.

[‡] IMPA, Instituto de Matemática Pura e Aplicada, Estrada Dona Castorina 110, Rio de Janeiro, RJ, CEP 22460-320, Brazil (benar@impa.br). This author's research was partially supported by CNPq grants 300755/2005-8 and 475647/2006-8, as well as by PRONEX-Optimization.

where $\lambda > 0$ is fixed, $\{\rho_k\} \subset [0,2]$ is a sequence of scalars, and, for any operator T, $R_{\lambda T}$ denotes the map $2(I+\lambda T)^{-1} - I$. References to this class of methods in the context of set-valued monotone operators include [16, 15, 9, 10]. The method of partial inverses [27] is a special case of this approach. If any solutions to $0 \in T_1(x) + T_2(x)$ exist, this method can be shown to converge weakly under mild assumptions on $\{\rho_k\}$ to some point $s = x + \lambda y$, where $y \in T_2(x)$ and $-y \in T_1(x)$, so $0 \in T_1(x) + T_2(x)$.

Another family of $n = 2$ methods is the double-backward class; see, for example, [17, 21, 7, 4]. These methods use the comparatively simple iteration

$$x^{k+1} = (I + \lambda_k T_1)^{-1}(I + \lambda_k T_2)^{-1}\left(x^k\right), \tag{1.3}$$

where $\{\lambda_k\}$ is a sequence of positive scalars. These methods have attractive convergence properties, but, unfortunately, solutions to (1.1) are not fixed points of (1.3) for general maximal monotone operators T_1 and T_2. Only if $\{\lambda_k\}$ approaches zero in a particular way is this approach known to solve (1.1). It does not appear that double-backward algorithms are used in practice for (1.1).

Splitting methods of the forward-backward class, generalizing standard gradient projection methods for variational inequalities and optimization problems, are more popular than double-backward methods and use the recursion

$$x^{k+1} \in (I + \lambda_k T_2)^{-1}\left(x^k - \lambda_k T_1\left(x^k\right)\right), \tag{1.4}$$

where $\{\lambda_k\}$ is again a sequence of positive scalars. References applying such methods to problems in the form (1.1) with $n = 2$ include [12, 28]. However, such methods must typically impose additional assumptions on at least one of the operators T_1 or T_2, usually T_1.

Traditionally, splitting algorithms allowing $n > 2$ have either explicitly or implicitly relied on reduction of (1.1) to the case $n = 2$ in the product space $\mathcal{H}^n$, endowed with the canonical inner product $\langle (x_1,\dots,x_n),(y_1,\dots,y_n)\rangle = \sum_{i=1}^n \langle x_i, y_i\rangle$, as follows: define the closed subspace

$$W \overset{\text{def}}{=} \{(w_1,\dots,w_n) \in \mathcal{H}^n \mid w_1 + w_2 + \cdots + w_n = 0\}, \tag{1.5}$$

whose orthogonal complement is

$$W^\perp = \{(v_1,\dots,v_n) \in \mathcal{H}^n \mid v_1 = v_2 = \cdots = v_n\} = \{(v,v,\dots,v) \mid v \in \mathcal{H}\}.$$

Next, define two operators $A, B : \mathcal{H}^n \rightrightarrows \mathcal{H}^n$ via $A \overset{\text{def}}{=} T_1 \otimes T_2 \otimes \cdots \otimes T_n$ and $B \overset{\text{def}}{=} N_{W^\perp}$, the normal cone map of $W^\perp$, that is,

$$A(x_1,\dots,x_n) = T_1(x_1) \times T_2(x_2) \times \cdots \times T_n(x_n), \tag{1.6}$$

$$B(x_1,\dots,x_n) = \begin{cases} W & \text{if } x_1 = x_2 = \cdots = x_n, \\ \emptyset & \text{otherwise.} \end{cases} \tag{1.7}$$

Using the maximal monotonicity of $T_1,\dots,T_n$, it is straightforward to establish that A and B are maximal monotone on $\mathcal{H}^n$ and that

$$\begin{aligned} & 0 \in A(x_1,\dots,x_n) + B(x_1,\dots,x_n) \\ \Leftrightarrow\quad & x_1 = x_2 = \cdots = x_n,\ \exists\, y_i \in T_i(x_i),\ i = 1,\dots,n : y_1 + y_2 + \cdots + y_n = 0 \\ \Leftrightarrow\quad & x_1 = x_2 = \cdots = x_n \text{ solves (1.1).} \end{aligned} \tag{1.8}$$

Applying Douglas–Rachford splitting to (1.8) produces Spingarn's method [27, section 5], in which one performs independent proximal steps on each of the operators $T_1, \ldots, T_n$ and then computes the next iterate by essentially averaging the results. In this setting, a proximal step on one operator cannot "feed" information into the proximal step for another operator within the same iteration. Applying a different $n = 2$ splitting method to (1.8) cannot alter this situation: evaluating the resolvent of A as defined in (1.6) will always yield independent, essentially simultaneous resolvent evaluations for $T_1, \ldots, T_n$.

A notable special case of (1.1) is the *convex feasibility* problem of finding a point in the intersection of n closed convex sets $C_1, \ldots, C_n$. If for $i = 1, \ldots, n$ one lets $T_i = N_{C_i}$, the normal cone map of the set C_i, then solving problem (1.1) is equivalent to finding a point $x \in \bigcap_{i=1}^n C_i$, and each resolvent operator $(I + \lambda T_i)^{-1}$ is simply the projection map onto the corresponding set C_i. In this special case, any solution x to (1.1) has the property of also being a zero of every constituent operator T_i, that is, if x satisfies $0 \in T_1(x) + \cdots + T_n(x)$, then one in fact has $0 \in T_i(x)$, $i = 1, \ldots, n$. This special property makes possible a wide variety of specialized methods composed of projection operations onto the individual sets C_i in a very flexible manner; see, for example, [2, 3] for some examples of the rich and extensive literature of such algorithms. Unfortunately, however, these algorithms do not generalize readily to the arbitrary maximal monotone operator setting (1.1), in which the solution x may not be a root of any of the constituent operators T_i. Thus, until quite recently [11], forward-backward, double-backward, and Peaceman/Douglas–Rachford, all using reduction to the case $n = 2$, were in essence the only splitting algorithms available for the general problem (1.1).

Here, we propose to take a new, projective approach to splitting algorithms for the general problem (1.1) with $n \geq 2$, generalizing our prior work [11] for the case $n = 2$. We make use of a product space, but in a somewhat different manner than the standard reduction to $n = 2$; instead, we define an *extended solution set* corresponding to (1.1) in the product space $\mathcal{H}^{n+1}$. We also employ projection, but not by trying to generalize successive projection methods for the special case $T_i = N_{C_i}$. Instead, we use a simple generic projection algorithm based on separating hyperplanes to produce a sequence weakly convergent to a point in the extended solution set. The decomposition properties of the algorithm arise from the particular way in which we construct the separating hyperplanes. This approach allows for a generality and flexibility not present in prior splitting methods for (1.1), while still applying in the general case.

The remainder of this paper is organized as follows: section 2 defines the extended solution set and analyzes some of its fundamental properties. To clarify the basic structure of our algorithm, we then introduce it in two stages: section 3 first describes a generic, abstract family of projection methods for finding a point in the extended solution set, giving general convergence conditions. Section 4 then specializes this abstract family to a concrete family characterized by a large number of parameters, presenting conditions under which it conforms to section 3's convergence conditions. Section 5 describes some variations and special cases of the algorithm of section 4, in particular showing that it subsumes Spingarn's method [27]. Section 6 describes some simple, preliminary computational experiments suggesting that our approach has the potential to converge significantly faster than prior splitting algorithms. Finally, section 7 gives some conclusions and topics for future research, while two appendices prove some technical results needed for sections 3 and 4.

2. The extended solution set and its separators. Consider now the Hilbert space $\mathcal{H} \times \mathcal{H}^n = \mathcal{H}^{n+1}$ under the canonical inner product

$$\langle (v, w_1, \dots, w_n), (x, y_1, \dots, y_n) \rangle = \langle v, x \rangle + \sum_{i=1}^{n} \langle w_i, y_i \rangle,$$

and define the closed linear subspace

$$V \overset{\text{def}}{=} \mathcal{H} \times W = \left\{ (v, w_1, \dots, w_n) \in \mathcal{H}^{n+1} \mid w_1 + \dots + w_n = 0 \right\}. \tag{2.1}$$

We define the *extended solution set* for problem (1.1) to be

$$S_{\mathrm{e}}(T_1, \dots, T_n) \overset{\text{def}}{=} \{ (z, w_1, \dots, w_n) \in V \mid w_i \in T_i(z),\ i = 1, \dots, n \}. \tag{2.2}$$

LEMMA 2.1. *Finding a point in* $S_{\mathrm{e}}(T_1, \dots, T_n)$ *is equivalent to solving* (1.1) *in the sense that*

$$0 \in T_1(z) + \dots + T_n(z) \iff \exists w_1, \dots, w_n \in \mathcal{H} : \ (z, w_1, \dots, w_n) \in S_{\mathrm{e}}(T_1, \dots, T_n).$$

Proof. For any $z \in \mathcal{H}$, $0 \in T_1(z) + \dots + T_n(z)$ if and only if there exist $w_1, \dots, w_n \in \mathcal{H}$ such that $w_1 + \dots + w_n = 0$ and $w_i \in T_i(z)$ for $i = 1, \dots, n$. This condition in turn holds if and only if $(z, w_1, \dots, w_n) \in S_{\mathrm{e}}(T_1, \dots, T_n)$. □

PROPOSITION 2.2. *If the monotone operators* $T_1, \dots, T_n : \mathcal{H} \rightrightarrows \mathcal{H}$ *are maximal, the corresponding extended solution set* $S_{\mathrm{e}}(T_1, \dots, T_n)$ *is closed and convex in* $\mathcal{H}^{n+1}$. *Further, if* $(z, w_1, \dots, w_n)$ *and* $(z', w_1', \dots, w_n')$ *are any two points in* $S_{\mathrm{e}}(T_1, \dots, T_n)$, *then* $\langle z - z', w_j - w_j' \rangle = 0$ *for all* $j = 1, \dots, n$.

Proof. Closedness of $S_{\mathrm{e}}(T_1, \dots, T_n)$ follows immediately from (2.2), the closedness of the linear subspace V, and the closedness of the graphs of the maximal monotone operators $T_1, \dots, T_n$. To prove convexity, take any scalars $p, q \geq 0$, $p + q = 1$, and any

$$(z, w_1, \dots, w_n), (z', w_1', \dots, w_n') \in S_{\mathrm{e}}(T_1, \dots, T_n).$$

If we can establish that

$$p(z, w_1, \dots, w_n) + q(z', w_1', \dots, w_n') \in S_{\mathrm{e}}(T_1, \dots, T_n),$$

the proof of convexity will be complete. Since $V \supset S_{\mathrm{e}}(T_1, \dots, T_n)$ is a linear subspace, it is clear that $p(z, w_1, \dots, w_n) + q(z', w_1', \dots, w_n') \in V$. From (2.2), it thus remains only to show that for all $j = 1, \dots, n$,

$$p w_j + q w_j' \in T_j(pz + qz'). \tag{2.3}$$

To this end, fix any $j \in \{1, \dots, n\}$. By the monotonicity of the T_i, we have that $\langle z - z', w_i - w_i' \rangle \geq 0$ for all $i = 1, \dots, n$, and so

$$0 \leq \langle z - z', w_j - w_j' \rangle \leq \sum_{i=1}^{n} \langle z - z', w_i - w_i' \rangle = \left\langle z - z', \sum_{i=1}^{n} w_i - \sum_{i=1}^{n} w_i' \right\rangle.$$

Since $\sum_1^n w_i = 0$ and $\sum_1^n w_i' = 0$, we conclude that $\langle z - z', w_j - w_j' \rangle = 0$, establishing the second statement of the proposition. Now, consider an arbitrary

$(\hat{z}, \hat{w}_j) \in \operatorname{graph}(T_j)$, and observe that

$$\begin{aligned}
&\left\langle \hat{z} - (pz + qz'), \hat{w}_j - \big(pw_j + qw'_j\big)\right\rangle \\
&= p\left\langle \hat{z} - z, \hat{w}_j - \big(pw_j + qw'_j\big)\right\rangle + q\left\langle \hat{z} - z', \hat{w}_j - \big(pw_j + qw'_j\big)\right\rangle \\
&= p\left\langle \hat{z} - z, \hat{w}_j - (1-q)w_j - qw'_j\right\rangle + q\left\langle \hat{z} - z', \hat{w}_j - pw_j - (1-p)w'_j\right\rangle \\
&= p\left\langle \hat{z} - z, \hat{w}_j - w_j\right\rangle + pq\left\langle \hat{z} - z, w_j - w'_j\right\rangle + q\left\langle \hat{z} - z', \hat{w}_j - w'_j\right\rangle + pq\left\langle \hat{z} - z', w'_j - w_j\right\rangle \\
&= p\left\langle \hat{z} - z, \hat{w}_j - w_j\right\rangle + q\left\langle \hat{z} - z', \hat{w}_j - w'_j\right\rangle + pq\left\langle (\hat{z} - z) - (\hat{z} - z'), w_j - w'_j\right\rangle \\
&= p\left\langle \hat{z} - z, \hat{w}_j - w_j\right\rangle + q\left\langle \hat{z} - z', \hat{w}_j - w'_j\right\rangle - pq\underbrace{\left\langle z - z', w_j - w'_j\right\rangle}_{=\,0} \\
&= p\left\langle \hat{z} - z, \hat{w}_j - w_j\right\rangle + q\left\langle \hat{z} - z', \hat{w}_j - w'_j\right\rangle.
\end{aligned}$$

The monotonicity of T_j implies that $\langle \hat{z} - z, \hat{w}_j - w_j\rangle \geq 0$ and $\langle \hat{z} - z', \hat{w}_j - w'_j\rangle \geq 0$, so we conclude that $\langle \hat{z} - (pz + qz'), \hat{w}_j - (pw_j + qw'_j)\rangle \geq 0$. Since $(\hat{z}, \hat{w}_j) \in \operatorname{graph}(T_j)$ was arbitrary and T_j is maximal, (2.3) holds, and $S_{\mathrm{e}}(T_1, \ldots, T_n)$ must be convex. □

Although we do not use it in the development of our algorithm, it is now nearly immediate to obtain the following result about $S_{\mathrm{e}}(T_1, \ldots, T_n)$.

COROLLARY 2.3. *The interior of* $S_{\mathrm{e}}(T_1, \ldots, T_n)$ *is empty relative to both* $\mathcal{H}^{n+1}$ *and* V.

Proof. Take any $p = (z, w_1, \ldots, w_n) \in S_{\mathrm{e}}(T_1, \ldots, T_n)$ and any nonzero $v \in \mathcal{H}$. For $\epsilon \geq 0$, consider the point $p'(\epsilon) = (z + \epsilon v, w_1 + \epsilon v, w_2 - \epsilon v, w_3, \ldots, w_n)$. As $p \in V$, we have $(w_1 + \epsilon v) + (w_2 - \epsilon v) + w_3 + \cdots + w_n = w_1 + \cdots + w_n = 0$, and so $p'(\epsilon) \in V$. Let int_V denote the interior relative to V. Now, if $p \in \operatorname{int}_V S_{\mathrm{e}}(T_1, \ldots, T_n)$, then we should have to have $p'(\epsilon) \in S_{\mathrm{e}}(T_1, \ldots, T_n)$ for all sufficiently small $\epsilon > 0$. But if $p'(\epsilon) \in S_{\mathrm{e}}(T_1, \ldots, T_n)$, then Proposition 2.2 implies $0 = \langle z - (z + \epsilon v), w_1 - (w_1 + \epsilon v)\rangle = \langle -\epsilon v, -\epsilon v\rangle = \epsilon^2\|v\|^2$. Since $v \neq 0$, we immediately obtain $\epsilon = 0$ and thus a contradiction. We conclude $p \notin \operatorname{int}_V S_{\mathrm{e}}(T_1, \ldots, T_n)$ and, since p was arbitrary, that $\operatorname{int}_V S_{\mathrm{e}}(T_1, \ldots, T_n) = \emptyset$. Since $S_{\mathrm{e}}(T_1, \ldots, T_n)$'s interior is empty relative to V, it is also empty relative to $\mathcal{H}^{n+1} \supset V$. □

Several variations on the definition of $S_{\mathrm{e}}(T_1, \ldots, T_n)$ are also possible. One possibility is to implicitly define w_n in terms of $w_1, \ldots, w_{n-1}$, obtaining

$$\left\{(z, w_1, \ldots, w_{n-1}) \;\middle|\; w_i \in T_i(z), i = 1, \ldots, n-1,\; -(w_1 + \cdots + w_{n-1}) \in T_n(z)\right\}.$$

This variation, in the case $n = 2$, is used in our earlier work [11]. Another possible variation is to use the set

$$\left\{ (z_1, \ldots, z_n, w_1, \ldots, w_n) \;\middle|\; \begin{array}{l} z_1 = z_2 = \cdots = z_n \\ w_1 + w_2 + \cdots + w_n = 0 \\ w_i \in T_i(z_i),\ i = 1, \ldots, n \end{array} \right\}, \tag{2.4}$$

which is the intersection of the sets $\operatorname{graph}(N_{W^\perp})$ and (after some permutation of indices) $\operatorname{graph}(T_1) \times \operatorname{graph}(T_2) \times \cdots \times \operatorname{graph}(T_n)$ in $\mathcal{H}^{2n}$. Such variations in the definition of $S_{\mathrm{e}}(T_1, \ldots, T_n)$ do not lead to material differences in the algorithms resulting from the analysis in sections 3 and 4 below.

In view of Lemma 2.1 and Proposition 2.2, we attempt to solve (1.1) by finding a point in $S_{\mathrm{e}}(T_1, \ldots, T_n)$, a problem we in turn approach by using a separator-projection algorithm. The separating hyperplanes used in our algorithm are constructed in a simple manner from points $(x_i, y_i) \in \operatorname{graph}(T_i)$, $i = 1, \ldots, n$. The following lemma details the construction and properties of these separators.

LEMMA 2.4. *Given $(x_i, y_i) \in \operatorname{graph}(T_i)$, $i = 1, \ldots, n$, define $\varphi : V \to \mathbb{R}$ via*

$$\varphi(z, w_1, \ldots, w_n) \overset{\text{def}}{=} \sum_{i=1}^{n} \langle z - x_i, y_i - w_i \rangle. \tag{2.5}$$

Then, for any $(z, w_1, \ldots, w_n) \in S_{\mathrm{e}}(T_1, \ldots, T_n)$, one has $\varphi(z, w_1, \ldots, w_n) \le 0$, that is,

$$S_{\mathrm{e}}(T_1, \ldots, T_n) \subseteq \{(z, w_1, \ldots, w_n) \in V \mid \varphi(z, w_1, \ldots, w_n) \le 0\}.$$

Additionally, φ is affine on V, with

$$\nabla\varphi = \left(\sum_{i=1}^{n} y_i, x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}\right), \qquad \textit{where} \qquad \bar{x} \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{2.6}$$

and

$$\begin{aligned} \nabla\varphi = 0 \quad &\Longleftrightarrow \quad (x_1, y_1, \ldots, y_n) \in S_{\mathrm{e}}(T_1, \ldots, T_n),\ x_1 = x_2 = \cdots = x_n \\ &\Longleftrightarrow \quad \varphi(z, w_1, \ldots, w_n) = 0 \ \ \forall (z, w_1, \ldots, w_n) \in V. \end{aligned}$$

Proof. Take any $(z, w_1, \ldots, w_n) \in S_{\mathrm{e}}(T_1, \ldots, T_n)$. For each $i = 1, \ldots, n$, we have $w_i \in T_i(z)$ and $y_i \in T(x_i)$. Since T_i is monotone, we have $\langle z - x_i, w_i - y_i \rangle \ge 0$. Negating and summing these inequalities, we conclude that $\varphi(z, w_1, \ldots, w_n) \le 0$, proving the first claim.

Next, take $(z, w_1, \ldots, w_n)$ to be an arbitrary element of V. Expanding and regrouping the inner products in (2.5), we obtain

$$\begin{aligned} \varphi(z, w_1, \ldots, w_n) &= \left\langle z, \sum_{i=1}^{n} y_i \right\rangle - \left\langle z, \sum_{i=1}^{n} w_i \right\rangle - \sum_{i=1}^{n} \langle x_i, y_i \rangle + \sum_{i=1}^{n} \langle x_i, w_i \rangle \qquad (2.7) \\ &= \left\langle z, \sum_{i=1}^{n} y_i \right\rangle - \sum_{i=1}^{n} \langle x_i, y_i \rangle + \sum_{i=1}^{n} \langle x_i, w_i \rangle \qquad (2.8) \\ &= \left\langle z, \sum_{i=1}^{n} y_i \right\rangle - \sum_{i=1}^{n} \langle x_i, y_i \rangle + \sum_{i=1}^{n} \langle x_i - \bar{x}, w_i \rangle + \left\langle \bar{x}, \sum_{i=1}^{n} w_i \right\rangle \\ &= \left\langle z, \sum_{i=1}^{n} y_i \right\rangle + \sum_{i=1}^{n} \langle x_i - \bar{x}, w_i \rangle - \sum_{i=1}^{n} \langle x_i, y_i \rangle \qquad (2.9) \\ &= \left\langle (z, w_1, \ldots, w_n), \left(\sum_{i=1}^{n} y_i, x_1 - \bar{x}, \ldots, x_n - \bar{x}\right) \right\rangle - \sum_{i=1}^{n} \langle x_i, y_i \rangle, \end{aligned}$$

where (2.8) and (2.9) follow from $\sum_{i=1}^{n} w_i = 0$ since $(z, w_1, \ldots, w_n) \in V$. Because $\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0$, we have that $(\sum_{i=1}^{n} y_i, x_1 - \bar{x}, \ldots, x_n - \bar{x}) \in V$. Thus, the last form of $\varphi(z, w_1, \ldots, w_n)$ above shows that φ is indeed an affine function on the space V, and $\nabla\varphi = (\sum_{i=1}^{n} y_i, x_1 - \bar{x}, \ldots, x_n - \bar{x})$.

Finally, we note that $\nabla\varphi = 0$ if and only if $\sum_{i=1}^{n} y_i = 0$ and $x_1 = \cdots = x_n = \bar{x}$. In that case, since $y_i \in T_i(x_i)$, $i = 1, \ldots, n$, one also has $(x_1, y_1, \ldots, y_n) = (\bar{x}, y_1, \ldots, y_n) \in S_{\mathrm{e}}(T_1, \ldots, T_n)$. In this case, we have $\sum_{i=1}^{n} \langle x_i, y_i \rangle = \langle x_1, \sum_{i=1}^{n} y_i \rangle = \langle x_1, 0 \rangle = 0$, and we conclude that φ is the zero function. □

Note that φ is not an affine function on the space $\mathcal{H}^{n+1}$ but only on its subspace V, where the "cross term" $\langle z, \sum_{i=1}^{n} w_i \rangle$ in (2.7) must be zero. We will thus implement our algorithm within the subspace V.

Next, it is natural to ask, given a point $(z, w_1, \ldots, w_n)$ in $V \backslash S_{\mathrm{e}}(T_1, \ldots, T_n)$, how to choose the pairs $(x_i, y_i) \in \operatorname{graph}(T_i)$ so that φ separates $(z, w_1, \ldots, w_n)$ from $S_{\mathrm{e}}(T_1, \ldots, T_n)$, that is, $\varphi(z, w_1, \ldots, w_n) > 0$. In fact, such a choice may be accomplished by a "prox" operation on each of the operators $T_1, \ldots, T_n$. By the maximal monotonicity of the T_i and the classic results of [18], there exists for each $i = 1, \ldots, n$ a unique $(x_i, y_i) \in \operatorname{graph}(T_i)$ such that $x_i + y_i = z + w_i$. Rearranging this equation, we obtain $z - x_i = y_i - w_i$ and thus that $\varphi(z, w_1, \ldots, w_n) = \sum_{i=1}^{n} \|z - x_i\|^2$. Thus, $\varphi(z, w_1, \ldots, w_n) > 0$ unless $x_1 = \cdots = x_n = z$, in which case it is easily deduced that $y_i = w_i$ for all i, and therefore $(z, w_1, \ldots, w_n) \in S_{\mathrm{e}}(T_1, \ldots, T_n)$, contrary to the assumption. Finding the necessary $(x_i, y_i) \in \operatorname{graph}(T_i)$ is equivalent to evaluating the resolvent $(I + T_i)^{-1}$, which is, by assumption, tractable for each individual T_i. We will greatly generalize this procedure for determining a separator in section 4 below.

3. An abstract family of projection algorithms.

We now have the necessary ingredients for implementing a projection method: a closed convex set S and at least one tractable procedure for calculating a separator between S and any $p \notin S$. Therefore, we may apply the following algorithmic template.

ALGORITHM 1. *Suppose $S \neq \emptyset$ is a closed convex set in a real Hilbert space U. Start with an arbitrary $p^0 \in U$. Then, for $k = 0, 1, \ldots,$ repeat the following:*

1. *Determine a non-constant differentiable affine function $\varphi_k : U \to \mathbb{R}$ such that $\varphi_k(p) \leq 0$ for all $p \in S$.*
2. *Let $\overline{p}^k$ be the projection of p^k onto the halfspace $H_k \overset{\mathrm{def}}{=} \{p \in U \mid \varphi_k(p) \leq 0\}$, that is,*

$$\overline{p}^k = p^k - \frac{\max\left\{0, \varphi_k\left(p^k\right)\right\}}{\|\nabla\varphi_k\|^2} \nabla\varphi_k. \tag{3.1}$$

3. *Choose some relaxation parameter $\rho_k \in (0, 2)$, and set*

$$p^{k+1} = p^k + \rho_k\left(\overline{p}^k - p^k\right).$$

The last two steps may simply be condensed to

$$p^{k+1} = p^k - \rho_k \frac{\max\left\{0, \varphi_k\left(p^k\right)\right\}}{\|\nabla\varphi_k\|^2} \nabla\varphi_k. \tag{3.2}$$

The basic properties of this algorithmic form may be derived by following the analysis of classical projection algorithms, dating back to Cimmino [6] and Kaczmarz [13, 14] in the late 1930s. A comprehensive review of projection algorithms may be found in [2]. As in any (relaxed) projection method, the sequences generated by Algorithm 1 behave as follows: for any $p^* \in S$, then $p^* \in H_k$ and the firm nonexpansiveness property of the projection mapping onto H_k assures that for all $k \geq 0$,

$$\left\|p^* - \overline{p}^k\right\|^2 \leq \left\|p^* - p^k\right\|^2 - \left\|\overline{p}^k - p^k\right\|^2, \tag{3.3}$$

$$\left\|p^* - p^{k+1}\right\|^2 \leq \left\|p^* - p^k\right\|^2 - \rho_k(2 - \rho_k)\left\|p^{k+1} - p^k\right\|^2. \tag{3.4}$$

The basic behavior of this class of methods is as follows; we omit the proof, which is entirely standard.

PROPOSITION 3.1. *Any infinite sequence $\{p^k\}$ generated by Algorithm* 1 *behaves as follows:*

1. *For any $p^* \in S$, the sequence $\{\|p^k - p^*\|\}$ is nonincreasing—that is, $\{p^k\}$ is Fejér monotone to S.*
2. *If $p^{k_0} \in S$ for some $k_0 \geq 0$, then $p^k = p^{k_0}$ for all $k \geq k_0$.*
3. *If $\{p^k\}$ has a strong accumulation point in S, then the whole sequence converges to that point.*
4. *If S is nonempty, then $\{p^k\}$ is bounded. Moreover, if there exist $\underline{\rho}, \overline{\rho}$ such that $0 < \underline{\rho} \leq \rho_k \leq \overline{\rho} < 2$ for all k, then*

$$\sum_{k=0}^{\infty} \left\|p^k - \overline{p}^k\right\|^2 < \infty, \qquad \sum_{k=0}^{\infty} \left\|p^k - p^{k+1}\right\|^2 < \infty. \tag{3.5}$$

5. *The sequence $\{p^k\}$ has at most one weak accumulation point in S.*

Note, however, that the basic template of Algorithm 1 is not sufficient to ensure weak convergence of $\{p^k\}$ to a point in S because the separators φ_k might not be chosen to actually separate p^k from S or might separate in a pathologically "shallow" way. The analysis of [11] guarantees convergence using the condition $\varphi_k(p^k) \geq \xi \left\|\nabla\varphi_k\right\|^2$ for all $k \geq 0$, where $\xi > 0$ is a fixed constant. We will also use this condition below.

We now restate and specialize Algorithm 1 for the case $S = S_{\text{e}}(T_1, \ldots, T_n)$ and $U = V$, with the separators constructed as in Lemma 2.4. We do not for the moment assume any particular way of choosing the $(x_i, y_i) \in \operatorname{graph}(T_i)$ yielding the separator.

ALGORITHM 2. *Start with an arbitrary $p^0 = (z^0, w_1^0, \ldots, w_n^0) \in V$. Then, for $k = 0, 1, \ldots,$ repeat the following:*

1. *For $i = 1, \ldots, n$, choose some $(x_i^k, y_i^k) \in \operatorname{graph}(T_i)$.*
2. *If $x_1^k = x_2^k = \cdots = x_n^k$ and $\sum_{i=1}^n y_i^k = 0$, let $w_i^{k+1} = y_i^k$ for $i = 1, \ldots, n$ and $z^{k+1} = x_1^k$. Otherwise, continue:*
3. *Let $\varphi_k : V \to \mathbb{R}$ be the separator derived from (x_i^k, y_i^k) via (2.5), that is,*

$$\varphi_k(z, w_1, \ldots, w_n) \overset{\text{def}}{=} \sum_{i=1}^{n} \left\langle z - x_i^k, y_i^k - w_i \right\rangle, \tag{3.6}$$

and let $p^{k+1} = (z^{k+1}, w_1^{k+1}, \ldots, w_n^{k+1})$ be the projection of p^k onto the half-space $H_k \overset{\text{def}}{=} \{p \in V \mid \varphi_k(p) \leq 0\}$, with an overrelaxation factor $\rho_k \in (0, 2)$, that is,

$$\bar{x}^k = \frac{1}{n}\sum_{i=1}^{n} x_i^k, \tag{3.7}$$

$$\theta_k = \frac{\max\left\{0, \sum_{i=1}^n \left\langle z^k - x_i^k, y_i^k - w_i^k\right\rangle\right\}}{\left\|\sum_{i=1}^n y_i^k\right\|^2 + \sum_{i=1}^n \left\|x_i^k - \bar{x}^k\right\|^2}, \tag{3.8}$$

$$z^{k+1} = z^k - \rho_k\theta_k\sum_{i=1}^{n} y_i^k, \tag{3.9}$$

$$w_i^{k+1} = w_i^k - \rho_k\theta_k\left(x_i^k - \bar{x}^k\right), \qquad i = 1, \ldots, n. \tag{3.10}$$

Note that the test in step 2 guarantees that the denominator in (3.8) cannot be zero. Formula (2.6) for the gradient of φ_k implies that (3.7)–(3.10) indeed calculate the overrelaxed projection of p^k onto H_k, and Algorithm 2 is thus Algorithm 1 specialized to $U = V$ and $S = S_{\text{e}}(T_1, \ldots, T_n)$. Note also that $p^k \in V$ and the update (3.10) ensure $w_1^{k+1} + \cdots + w_n^{k+1} = 0$, so by induction the entire iterate sequence $\{p^k\} = \{(z^k, w_1^k, \ldots, w_n^k)\}$ produced by Algorithm 2 lies in V.

We now perform a preliminary analysis of the convergence properties of Algorithm 2.

PROPOSITION 3.2. *Suppose that the following conditions are met in Algorithm* 2:

1. $S_{\mathrm{e}}(T_1,\ldots,T_n) \neq \emptyset$.
2. $0 < \underline{\rho} \leq \rho_k \leq \overline{\rho} < 2$ *for all* k.
3. *There exists some scalar* $\xi > 0$ *such that, for all* $k \geq 0$,

$$\text{(3.11)}\quad \varphi_k\big(p^k\big) = \varphi_k\big(t(z^k, w_1^k, \ldots, w_n^k\big) \geq \xi \left\|\nabla\varphi_k\right\|^2 = \xi\left(\left\|\sum_{i=1}^n y_i^k\right\|^2 + \sum_{i=1}^n \left\|x_i^k - \bar{x}^k\right\|^2\right).$$

Then $\nabla\varphi_k \to 0$, *that is,* $x_i^k - x_j^k \to 0$ *for all* $i,j = 1,\ldots,n$, *and* $\sum_{i=1}^n y_i^k \to 0$. *Furthermore,* $\varphi_k(p^k) \to 0$. *If it is also true that*

4. *either* $\mathcal{H}$ *has finite dimension or the operator* $T_1 + \cdots + T_n$ *is maximal,*
5. $z^k - \bar{x}^k \to 0$,
6. $w_i^k - y_i^k \to 0$, *for* $i = 1,\ldots,n$,

then $\{p^k\}$ *converges weakly to some* $p^\infty = (z^\infty, w_1^\infty, \ldots, w_n^\infty) \in S_{\mathrm{e}}(T_1,\ldots,T_n)$, *which implies that* z^∞ *solves* (1.1). *Furthermore,* $x_i^k \xrightarrow{\mathsf{w}} z^\infty$ *and* $y_i^k \xrightarrow{\mathsf{w}} w_i^\infty$ *for* $i = 1,\ldots,n$.

Proof. The hypothesis that $\varphi_k(p^k) \geq \xi\left\|\nabla\varphi_k\right\|^2$ implies that $\varphi_k(p^k)$ is always nonnegative, so we obtain from (3.1) that

$$\text{(3.12)}\qquad \left\|p^k - \overline{p}^k\right\| = \varphi_k\left(p^k\right)/\left\|\nabla\varphi_k\right\|$$

for all k having $\nabla\varphi_k \neq 0$. Substituting $\varphi_k(p^k) \geq \xi\left\|\nabla\varphi_k\right\|^2$ into this equation, we obtain

$$\text{(3.13)}\qquad \left\|p^k - \overline{p}^k\right\| \geq \xi\left\|\nabla\varphi_k\right\|,$$

which clearly also holds for k having $\nabla\varphi_k = 0$. From (3.5), which must hold by hypotheses 1–2 and Proposition 3.1(4), we have $\|p^k - \overline{p}^k\| \to 0$, so (3.13) implies $\nabla\varphi_k \to 0$. From the expression for $\nabla\varphi_k$ in (2.6), we immediately have $\sum_{i=1}^n y_i^k \to 0$ and $x_i^k - \bar{x}^k \to 0$ for $i = 1,\ldots,n$, and thus $x_i^k - x_j^k \to 0$ for all $i,j = 1,\ldots,n$. Multiplying (3.12) by $\|\nabla\varphi_k\|$, we obtain that

$$\text{(3.14)}\qquad \varphi_k\left(p^k\right) = \left\|p^k - \overline{p}^k\right\|\left\|\nabla\varphi_k\right\|$$

whenever $\nabla\varphi_k \neq 0$. By Lemma 2.4, $\varphi_k(p^k) = 0$ whenever $\nabla\varphi_k = 0$, so (3.14) holds for all $k \geq 0$. Since we have established that $\nabla\varphi_k \to 0$ and we also know that $\|p^k - \overline{p}^k\| \to 0$, (3.14) implies that $\varphi_k(p^k) \to 0$. The first set of conclusions are now established; note also that by hypothesis 1 and Fejér monotonicity, the sequence $\{p^k\}$ is bounded.

To prove the remainder of the proposition, we now assume that hyphotheses 4–6 also hold. From hypothesis 5 and $x_i^k - \bar{x}^k \to 0$, we immediately obtain

$$\text{(3.15)}\qquad z^k - x_i^k \to 0, \qquad i = 1,\ldots,n.$$

In hypothesis 4, suppose first that $\mathcal{H}$ is finite-dimensional. Consider any cluster point $p^\infty = (z^\infty, w_1^\infty, \ldots, w_n^\infty)$ of the bounded sequence $\{p^k\}$. There then exists a subsequence $\{p^k\}_{k\in\mathcal{K}}$ converging to p^∞. From (3.15), we then have $x_i^k \to_{k\in\mathcal{K}} z^\infty$, $i = 1,\ldots,n$. Similarly, hypothesis 6 implies $y_i^k \to_{k\in\mathcal{K}} w_i^\infty$, $i = 1,\ldots,n$. Because

$(x_i^k, y_i^k) \in \operatorname{graph}(T_i)$ for all i and k, and the maximality of the operators T_i implies that the sets $\operatorname{graph}(T_i)$ are closed, we then obtain $w_i^\infty \in T_i(z^\infty)$ for all $i = 1, \ldots, n$. Furthermore, since $\{p^k\} \subset V$ and V is a closed subspace, we also have $p^\infty \in V$ and thus $p^\infty \in S_{\mathrm{e}}(T_1, \ldots, T_n)$. Finally, we apply Proposition 3.1(3) to obtain that the entire sequence $\{p^k\}$ converges to $p^\infty \in S_{\mathrm{e}}(T_1, \ldots, T_n)$.

Now, assume the other alternative in hypothesis 4, that $T_1 + \cdots + T_n$ is maximal monotone. Let p^∞ be any weak cluster point of $\{p^k\}$. Then there exists a subsequence $\{p^k\}_{k \in \mathcal{K}}$ weakly convergent to p^∞, and, using hypotheses 5 and 6, we conclude that $(x_i^k, y_i^k) \xrightarrow{\mathsf{w}}_{k \in \mathcal{K}} (z^\infty, w_i^\infty)$, $i = 1, \ldots, n$. Next, we apply Proposition A.1 from Appendix A to conclude that $p^\infty = (z^\infty, w_1^\infty, \ldots, w_n^\infty) \in S_{\mathrm{e}}(T_1, \ldots, T_n)$. Since p^∞ was chosen arbitrarily, all weak cluster points of $\{p^k\}$ are in $S_{\mathrm{e}}(T_1, \ldots, T_n)$. Then we may apply Proposition 3.1(5) to conclude that the entire sequence $\{p^k\}$ converges weakly to p^∞.

In either case, the remaining conclusions follow from hypothesis 6 and (3.15). □

4. A general projective splitting scheme. To convert Algorithm 2 into an implementable procedure for solving (1.1), we must specify a way of choosing the $(x_i^k, y_i^k) \in \operatorname{graph}(T_i)$ so that the hypotheses of Proposition 3.2 are satisfied. One simple approach, as mentioned at the end of section 2, would be to choose the unique $(x_i^k, y_i^k) \in \operatorname{graph}(T_i)$ satisfying $x_i^k + y_i^k = z^k + w_i^k$. A simple generalization would be to add a proximal parameter $\lambda_i^k > 0$, yielding

$$x_i^k + \lambda_i^k y_i^k = z^k + \lambda_i^k w_i^k. \tag{4.1}$$

This scheme may in fact be greatly generalized without sacrificing its basic decomposability. Suppose for the moment that in each iteration we perform the proximal calculations for the T_i sequentially, starting with $i = 1$ and finishing with $i = n$. We may then wish to use the "recent" information generated in calculating (x_j^k, y_j^k), where $j < i$, when calculating (x_i^k, y_i^k). Specifically, when calculating (x_i^k, y_i^k), we consider replacing z^k with an affine combination of the vectors z^k and x_j^k, $j < i$. In particular, we first find the unique $(x_1^k, y_1^k) \in \operatorname{graph}(T_1)$ such that

$$x_1^k + \lambda_1^k y_1^k = z^k + \lambda_1^k w_1^k.$$

We next take some $\alpha_{21}^k \in \mathbb{R}$ and find the unique $(x_2^k, y_2^k) \in \operatorname{graph}(T_2)$ such that

$$x_2^k + \lambda_2^k y_2^k = \left(1 - \alpha_{21}^k\right) z^k + \alpha_{21}^k x_1^k + \lambda_2^k w_2^k.$$

To continue, we choose some $\alpha_{31}^k, \alpha_{32}^k \in \mathbb{R}$ and find the unique $(x_3^k, y_3^k) \in \operatorname{graph}(T_3)$ such that

$$x_3^k + \lambda_3^k y_3^k = \left(1 - \alpha_{31}^k - \alpha_{32}^k\right) z^k + \alpha_{31}^k x_1^k + \alpha_{32}^k x_2^k + \lambda_3^k w_3^k$$

and so forth. In general, we choose (x_i^k, y_i^k) to satisfy the conditions

$$x_i^k + \lambda_i^k y_i^k = \left(1 - \sum_{j=1}^{i-1} \alpha_{ij}^k\right) z^k + \sum_{j=1}^{i-1} \alpha_{ij}^k x_j^k + \lambda_i^k w_i^k, \qquad y_i^k \in T_i\left(x_i^k\right). \tag{4.2}$$

In addition to the flexibility afforded by the choice of the α_{ij}^k and λ_i^k, we consider several further generalizations of (4.2):

- We will allow errors $e_i^k \in \mathcal{H}$ in satisfying (4.2), so long as they satisfy the approximation criterion (4.8) below.
- The order of processing the operators may vary from iteration to iteration. At iteration k, we process the operators in the order specified by an arbitrary permutation $\pi_k(\cdot)$ of $\{1, \ldots, n\}$.

Thus, we arrive at the general scheme that, for all $i = 1, \ldots, n$ and $k \geq 0$, we have $y_i^k \in T_i(x_i^k)$ and

$$x_{\pi_k(i)}^k + \lambda_i^k y_{\pi_k(i)}^k = \left(1 - \sum_{j=1}^{i-1} \alpha_{ij}^k\right) z^k + \sum_{j=1}^{i-1} \alpha_{ij}^k x_{\pi_k(j)}^k + \lambda_i^k w_{\pi_k(i)}^k + e_i^k. \tag{4.3}$$

Note that the notion of processing the operators in some particular order $\pi_k(\cdot)$ does not necessarily preclude parallelism over i in evaluating (4.3), depending on how one chooses the α_{ij}^k. For example, if we choose $\alpha_{ij}^k = 0$ for all i, j and set the error terms $e_i^k = 0$, then (4.3) reduces to (4.1), which may be calculated independently and in parallel over i.

To analyze this scheme, we will employ some standard matrix analysis: given an $n \times n$ real matrix $\mathbf{L}$, we define $\|\mathbf{L}\|$ to be its operator 2-norm and $\kappa(\mathbf{L})$ to be the smallest eigenvalue of its symmetric part, that is,

$$\|\mathbf{L}\| \stackrel{\text{def}}{=} \max_{\substack{\mathbf{x} \in \mathbb{R}^n \\ \|\mathbf{x}\| = 1}} \|\mathbf{L}\mathbf{x}\| \qquad \operatorname{sym} \mathbf{L} \stackrel{\text{def}}{=} \tfrac{1}{2}(\mathbf{L} + \mathbf{L}^\top), \qquad \kappa(\mathbf{L}), \stackrel{\text{def}}{=} \min \operatorname{eig} \operatorname{sym} \mathbf{L}.$$

Note that it is straightforward to show that $\kappa(\mathbf{L}) \leq \|\mathbf{L}\|$ and that, for any $\mathbf{x} \in \mathbb{R}^n$, $\langle \mathbf{x}, \mathbf{L}\mathbf{x} \rangle \geq \kappa(\mathbf{L}) \|\mathbf{x}\|^2$. Analogously to the usual linear map $\mathbb{R}^n \to \mathbb{R}^n$ associated with $\mathbf{L}$, we can define a linear mapping $\mathcal{H}^n \to \mathcal{H}^n$ corresponding to $\mathbf{L}$ via

$$\mathbf{L}u = \mathbf{L}(u_1, \ldots, u_n) = (v_1, \ldots, v_n), \quad \text{where} \quad v_i = \sum_{j=1}^{n} \ell_{ij} u_j \in \mathcal{H}, \tag{4.4}$$

with ℓ_{ij} denoting the elements of $\mathbf{L}$. As one would intuitively expect, this mapping retains key spectral properties that $\mathbf{L}$ exhibits over $\mathbb{R}^n$.

LEMMA 4.1. *Let $\mathbf{L}$ be any $n \times n$ real matrix. For all $u = (u_1, \ldots, u_n) \in \mathcal{H}^n$,*

$$\|\mathbf{L}u\| \leq \|\mathbf{L}\| \, \|u\|, \tag{4.5}$$

$$\langle u, \mathbf{L}u \rangle \geq \kappa(\mathbf{L}) \|u\|^2, \tag{4.6}$$

where $\mathbf{L}u$ is defined by (4.4), $\langle \cdot, \cdot \rangle$ denotes the canonical inner product for $\mathcal{H}^n$ induced by the inner product for $\mathcal{H}$, and $\|\cdot\|$ applied to elements of $\mathcal{H}^n$ denotes the norm induced by this inner product.

Appendix B proves this result. Of particular interest are the $n \times n$ matrices $\mathbf{\Lambda}_k = \operatorname{diag}(\lambda_1^k, \lambda_2^k, \ldots, \lambda_n^k)$ and $\mathbf{A}_k = [a_{ij}^{(k)}]_{i,j=1,\ldots,n}$, where

$$a_{ij}^{(k)} = \begin{cases} 1, & i = j, \\ -\alpha_{ij}^k, & i > j, \\ 0, & i < j. \end{cases}$$

We will show that if there exist $\beta, \zeta > 0$ such that

$$\kappa\left(\mathbf{\Lambda}_k^{-1} \mathbf{A}_k\right), \geq \zeta \qquad \left\|\mathbf{\Lambda}_k^{-1} \mathbf{A}_k\right\| \leq \beta \qquad \forall\, k \geq 0, \tag{4.7}$$

then choosing the $(x_i^k, y_i^k) \in \operatorname{graph}(T_i)$ via (4.3) will meet all of the hypotheses of Proposition 3.2, and we will obtain weak convergence. Stated in full, including the approximate calculation criterion, the algorithm is as follows.

Algorithm 3. *Choose scalars $\beta, \zeta > 0$, $0 < \underline{\rho} \le \overline{\rho} < 2$, and $\sigma \in [0,1)$. Start with an arbitrary $(z^0, w_1^0, \dots, w_n^0) \in V$, that is, $z^0, w_1^0, \dots, w_n^0 \in \mathcal{H}$ with $w_1^0 + \cdots + w_n^0 = 0$. Then, for $k = 0, 1, \dots,$ repeat the following:*

1. *Choose scalars $\lambda_i^k > 0$, $i = 1, \dots, n$, and α_{ij}^k, $1 \le j < i \le n$, such that $\kappa(\mathbf{\Lambda}_k^{-1}\mathbf{A}_k) \ge \zeta$ and $\|\mathbf{\Lambda}_k^{-1}\mathbf{A}_k\| \le \beta$, where $\mathbf{\Lambda}_k$ and $\mathbf{A}_k$ are defined as above. Let $\pi_k(\cdot)$ be any permutation of $\{1, \dots, n\}$. For $i = 1, \cdots, n$, find $(x_i^k, y_i^k) \in$* graph(T_i) *satisfying* (4.3) *and*

$$\sum_{i=1}^{n} \left(\lambda_i^k\right)^{-2} \left\|e_i^k\right\|^2 \le \sigma^2 \kappa\left(\mathbf{\Lambda}_k^{-1}\mathbf{A}_k\right)^2 \sum_{i=1}^{n} \left\|x_i^k - z^k\right\|^2. \tag{4.8}$$

2. *If $x_1^k = x_2^k = \cdots = x_n^k$ and $\sum_{i=1}^n y_i^k = 0$, let $w_i^{k+1} = y_i^k$ for $i = 1, \dots, n$ and $z^{k+1} = x_1^k$. Otherwise, continue:*
3. *Choose some $\rho_k \in [\underline{\rho}, \overline{\rho}]$, and set*

$$\bar{x}^k = \frac{1}{n}\sum_{i=1}^{n} x_i^k, \tag{4.9}$$

$$\theta_k = \frac{\sum_{i=1}^n \left\langle z^k - x_i^k, y_i^k - w_i^k \right\rangle}{\left\|\sum_{i=1}^n y_i^k\right\|^2 + \sum_{i=1}^n \left\|x_i^k - \bar{x}^k\right\|^2}, \tag{4.10}$$

$$z^{k+1} = z^k - \rho_k \theta_k \sum_{i=1}^{n} y_i^k, \tag{4.11}$$

$$w_i^{k+1} = w_i^k - \rho_k \theta_k \left(x_i^k - \bar{x}^k\right), \qquad i = 1, \dots, n. \tag{4.12}$$

The error condition (4.8) is an n-operator generalization of the relative error tolerance proposed in [24, 23, 26] for modified proximal-extragradient projection methods. Note that β's only role in the statement of the algorithm is to guarantee $\|\mathbf{\Lambda}_k^{-1}\mathbf{A}_k\|$ remains bounded, that is, that $\{\mathbf{\Lambda}_k^{-1}\mathbf{A}_k\}$ is a bounded sequence of matrices. Such boundedness may be assured by any sufficient condition bounding the absolute values $|a_{ij}^{(k)}/\lambda_i^k|$ of all entries of $\{\mathbf{\Lambda}_k^{-1}\mathbf{A}_k\}$. For example, if there exist $\underline{\lambda}, \overline{\alpha} \ge 0$ such that $\lambda_i^k \ge \underline{\lambda}$ and $|\alpha_{ij}^k| \le \overline{\alpha}$ for all $k \ge 1$, $i = 1, \dots, n$, and $1 \le j < i$, then $\{\mathbf{\Lambda}_k^{-1}\mathbf{A}_k\}$ must be bounded, and some β satisfying the condition $\|\mathbf{\Lambda}_k^{-1}\mathbf{A}_k\| \le \beta$ for all $k \ge 0$ must exist. In practice, we may therefore substitute conditions such as $\lambda_i^k \ge \underline{\lambda}$ and $|\alpha_{ij}^k| \le \overline{\alpha}$ for the condition $\|\mathbf{\Lambda}_k^{-1}\mathbf{A}_k\| \le \beta$ in step 3 of the algorithm. We now prove convergence of the method.

Proposition 4.2. *Suppose that either $\mathcal{H}$ has finite dimension or the operator $T_1 + \cdots + T_n$ is maximal. Suppose also that* (1.1) *has a solution. Then, in Algorithm* 3, *the sequences $\{z^k\}, \{x_1^k\}, \dots, \{x_n^k\} \subset \mathcal{H}$ all weakly converge to some z^∞ solving* (1.1). *For each $i = 1, \dots, n$, we also have $w_i^k, y_i^k \xrightarrow{\mathsf{w}} y_i^\infty$, where $y_i^\infty \in T_i(z^\infty)$ and also $y_1^\infty + \cdots + y_n^\infty = 0$.*

Proof. Define auxiliary sequences $\{p^k\} \subset \mathcal{H}^{n+1}$, $\{u^k\} = \{(u_1^k, \dots, u_n^k)\} \subset \mathcal{H}^n$, and $\{v^k\} = \{(v_1^k, \dots, v_n^k)\} \subset \mathcal{H}^n$ via

$$p^k \overset{\text{def}}{=} \left(z^k, w_1^k, \dots, w_n^k\right), \qquad u_i^k \overset{\text{def}}{=} x_i^k - z^k, \qquad v_i^k \overset{\text{def}}{=} w_i^k - y_i^k \tag{4.13}$$

for all $i = 1, \ldots, n$ and $k \geq 0$, and also define as in (3.6) the function

$$\varphi_k(p) = \varphi_k(z, w_1, \ldots, w_n) \overset{\text{def}}{=} \sum_{i=1}^{n} \left\langle z - x_i^k, y_i^k - w_i \right\rangle .$$

From (4.13), we immediately have

$$\begin{aligned} \varphi_k\left(p^k\right) = \varphi_k\left(z^k, w_1^k, \ldots, w_n^k\right) &= \sum_{i=1}^{n} \left\langle z^k - x_i^k, y_i^k - w_i^k \right\rangle \\ &= \left\langle u^k, v^k \right\rangle = \sum_{i=1}^{n} \left\langle u_i^k, v_i^k \right\rangle. \qquad (4.14) \end{aligned}$$

Further, define $e^k = (e_1^k, \ldots, e_n^k) \in \mathcal{H}^n$ for all $k \geq 0$, and observe that by taking square roots and substitution of the definitions of e^k and u^k, (4.8) simplifies via the notation (4.4) and the definition of $\mathbf{\Lambda}_k$ to

$$\left\| \mathbf{\Lambda}_k^{-1} e^k \right\| \leq \sigma \kappa\left(\mathbf{\Lambda}_k^{-1} \mathbf{A}_k\right) \left\| u^k \right\| . \qquad (4.15)$$

Take any $i \in \{1, \ldots, n\}$. Subtracting z^k from both sides of (4.3) and regrouping yields

$$\begin{aligned} &\left(x_{\pi_k(i)}^k - z^k\right) + \lambda_i^k y_{\pi_k(i)}^k = \sum_{j=1}^{i-1} \alpha_{ij}^k \left(x_{\pi_k(j)}^k - z^k\right) + \lambda_i^k w_{\pi_k(i)}^k + e_i^k \\ \Leftrightarrow \quad &\left(x_{\pi_k(i)}^k - z^k\right) - \sum_{j=1}^{i-1} \alpha_{ij}^k \left(x_{\pi_k(j)}^k - z^k\right) - e_i^k = \lambda_i^k \left(w_{\pi_k(i)}^k - y_{\pi_k(i)}^k\right) . \end{aligned}$$

Dividing by λ_i^k and substituting the definitions of u_i^k and v_i^k yields

$$\left(\frac{1}{\lambda_i^k}\right) \left(u_{\pi_k(i)}^k - \sum_{j=1}^{i-1} \alpha_{ij}^k u_{\pi_k(j)}^k - e_i^k \right) = v_{\pi_k(i)}^k . \qquad (4.16)$$

Applying the notation (4.4) to (4.16) for $i = 1, \ldots, n$ produces

$$v^k = \left(\mathbf{\Pi}_k \mathbf{\Lambda}_k^{-1} \mathbf{A}_k \mathbf{\Pi}_k^\top\right) u^k - \left(\mathbf{\Pi}_k \mathbf{\Lambda}_k^{-1}\right) e^k, \qquad (4.17)$$

where $\mathbf{\Pi}_k$ is the $n \times n$ permutation matrix corresponding to the permutation $\pi_k(\cdot)$. Substituting (4.17) into (4.14) yields

$$\begin{aligned} \varphi_k\left(p^k\right) &= \left\langle u^k, \mathbf{\Pi}_k \mathbf{\Lambda}_k^{-1} \mathbf{A}_k \mathbf{\Pi}_k^\top u^k \right\rangle - \left\langle u^k, \mathbf{\Pi}_k \mathbf{\Lambda}_k^{-1} e^k \right\rangle \\ &\geq \left\langle u^k, \mathbf{\Pi}_k \mathbf{\Lambda}_k^{-1} \mathbf{A}_k \mathbf{\Pi}_k^\top u^k \right\rangle - \left\| u^k \right\| \left\| \mathbf{\Pi}_k \mathbf{\Lambda}_k^{-1} e^k \right\| && [\text{Cauchy–Schwarz}] \\ &\geq \kappa\left(\mathbf{\Pi}_k \mathbf{\Lambda}_k^{-1} \mathbf{A}_k \mathbf{\Pi}_k^\top\right) \left\| u^k \right\|^2 - \left\| u^k \right\| \left\| \mathbf{\Pi}_k \mathbf{\Lambda}_k^{-1} e^k \right\| && [\text{using } (4.6)] \\ &= \kappa\left(\mathbf{\Lambda}_k^{-1} \mathbf{A}_k\right) \left\| u^k \right\|^2 - \left\| u^k \right\| \left\| \mathbf{\Lambda}_k^{-1} e^k \right\| && [\mathbf{\Pi}_k \text{ orthonormal}] \\ &\geq \kappa\left(\mathbf{\Lambda}_k^{-1} \mathbf{A}_k\right) \left\| u^k \right\|^2 - \sigma \kappa\left(\mathbf{\Lambda}_k^{-1} \mathbf{A}_k\right) \left\| u^k \right\|^2 && [\text{using } (4.15)] \\ &= (1 - \sigma) \kappa\left(\mathbf{\Lambda}_k^{-1} \mathbf{A}_k\right) \left\| u^k \right\|^2 \\ &\geq (1 - \sigma) \zeta \left\| u^k \right\|^2 . && [\text{using } (4.7)] \qquad (4.18) \end{aligned}$$

To meet hypothesis 3 of Proposition 3.2, we need to convert this lower bound on $\varphi_k(p^k)$, expressed in terms of $\|u^k\|^2$, to one expressed in terms of $\|\nabla\varphi_k\|^2$. To do so, first note that since $\sum_{i=1}^n w_i^k = 0$,

$$\sum_{i=1}^n v_i^k = \sum_{i=1}^n \left(w_i^k - y_i^k\right) = -\sum_{i=1}^n y_i^k \quad\Rightarrow\quad \left\|\sum_{i=1}^n v_i^k\right\|^2 = \left\|\sum_{i=1}^n y_i^k\right\|^2. \tag{4.19}$$

Next, define

$$\bar{u}^k \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n u_i^k = \frac{1}{n}\sum_{i=1}^n \left(x_i^k - z^k\right) = \bar{x}^k - z^k,$$

and observe that for all $i = 1, \dots, n$ and $k \geq 0$,

$$u_i^k - \bar{u}^k = x_i^k - z^k - \left(\bar{x}^k - z^k\right) = x_i^k - \bar{x}^k. \tag{4.20}$$

Substituting (4.19) and (4.20) into the expression for $\|\nabla\varphi_k\|^2$ arising from Lemma 2.4, we obtain

$$\begin{aligned}\|\nabla\varphi_k\|^2 &= \left\|\sum_{i=1}^n y_i^k\right\|^2 + \sum_{i=1}^n \left\|x_k^k - \bar{x}^k\right\|^2 \\ &= \left\|\sum_{i=1}^n v_i^k\right\|^2 + \sum_{i=1}^n \left\|u_i^k - \bar{u}^k\right\|^2 \\ &= \frac{1}{n}\left\|\mathbf{E}v^k\right\|^2 + \left\|\mathbf{M}u^k\right\|^2,\end{aligned}$$

where we define $\mathbf{E}$ to be the $n \times n$ matrix of all ones and $\mathbf{M} = \mathbf{I} - (1/n)\mathbf{E}$. Applying (4.5), it then follows that

$$\|\nabla\varphi_k\|^2 \leq \frac{1}{n}\left\|\mathbf{E}v^k\right\|^2 + \left\|\mathbf{M}u^k\right\|^2 \leq \frac{1}{n}\|\mathbf{E}\|^2\left\|v^k\right\|^2 + \|\mathbf{M}\|^2\left\|u^k\right\|^2. \tag{4.21}$$

Over $\mathbb{R}^n$, the matrix $\mathbf{M}$ represents orthogonal projection onto the nontrivial subspace $T = \{(t_1, \dots, t_n) \in \mathbb{R}^n \mid t_1 + \cdots + t_n = 0\}$, so we conclude $\|\mathbf{M}\| = 1$. It also follows that $I - M$ represents orthogonal projection onto the nontrivial subspace $T^\perp$, so $\|I - M\| = 1$ and $\|E\| = \|n(I - M)\| = n\,\|I - M\| = n$. Therefore, (4.21) reduces to

$$\|\nabla\varphi_k\|^2 \leq \left(\frac{1}{n}\right) n^2\left\|v^k\right\|^2 + \left\|u^k\right\|^2 = n\left\|v^k\right\|^2 + \left\|u^k\right\|^2. \tag{4.22}$$

Starting with (4.17), we obtain

$$\begin{aligned}\left\|v^k\right\|^2 &= \left\|\left(\mathbf{\Pi}_k\mathbf{\Lambda}_k^{-1}\mathbf{A}_k\mathbf{\Pi}_k^\top\right)u^k - \mathbf{\Pi}_k\mathbf{\Lambda}_k^{-1}e^k\right\|^2 \\ &\leq \left(\left\|\left(\mathbf{\Pi}_k\mathbf{\Lambda}_k^{-1}\mathbf{A}_k\mathbf{\Pi}_k^\top\right)u^k\right\| + \left\|\mathbf{\Pi}_k\mathbf{\Lambda}_k^{-1}e^k\right\|\right)^2 && [\,\text{triangle inequality}\,] \\ &\leq \left(\left\|\mathbf{\Pi}_k\mathbf{\Lambda}_k^{-1}\mathbf{A}_k\mathbf{\Pi}_k^\top\right\|\left\|u^k\right\| + \left\|\mathbf{\Lambda}_k^{-1}e^k\right\|\right)^2 && [\,\text{using (4.5)}\,] \\ &\leq \left(\left\|\mathbf{\Lambda}_k^{-1}\mathbf{A}_k\right\|\left\|u^k\right\| + \sigma\kappa\left(\mathbf{\Lambda}_k^{-1}\mathbf{A}_k\right)\left\|u^k\right\|\right)^2 && [\,\text{using (4.15)}\,] \\ &\leq \left((1+\sigma)\left\|\mathbf{\Lambda}_k^{-1}\mathbf{A}_k\right\|\left\|u^k\right\|\right)^2 && [\,\kappa(\mathbf{\Lambda}_k^{-1}\mathbf{A}_k) \leq \|\mathbf{\Lambda}_k^{-1}\mathbf{A}_k\|\,] \\ &\leq \left((1+\sigma)\beta\left\|u^k\right\|\right)^2 && [\,\text{using (4.7)}\,] \\ &= (1+\sigma)^2\beta^2\left\|u^k\right\|^2. && (4.23)\end{aligned}$$

Combining (4.22) and (4.23) yields

$$\|\nabla\varphi_k\|^2 \le \left(n(1+\sigma)^2\beta^2+1\right)\|u^k\|^2.$$

Combining this inequality with (4.18) yields

$$\varphi_k(p^k) \ge \frac{(1-\sigma)\zeta}{n(1+\sigma)^2\beta^2+1}\|\nabla\varphi_k\|^2, \tag{4.24}$$

implying that hypothesis 3 of Proposition 3.2 is satisfied by setting

$$\xi = \frac{(1-\sigma)\zeta}{n(1+\sigma)^2\beta^2+1} > 0.$$

Note that (4.24) implies $\varphi_k(p^k)$ is always nonnegative so that (4.10) is equivalent to (3.8), even though the $\max\{0,\cdot\}$ operation is omitted. In view of (4.24), Proposition 3.2 guarantees that $\nabla\varphi_k \to 0$ and $\varphi_k(p^k) \to 0$. From (4.18), we then conclude $u^k \to 0$, from which (4.23) implies that $v^k \to 0$. Thus, we have $u_i^k = x_i^k - z^k \to 0$ and $v_i^k = w_i^k - y_i^k \to 0$ for all $i = 1,\dots,n$, fulfilling hypotheses 5 and 6 of Proposition 3.2. Hypothesis 4 is satisfied by assumption, so all of the hypotheses of Proposition 3.2 hold. The (weak) convergence of the sequences $\{z^k\}$, $\{x_i^k\}$, $\{w_i^k\}$, and $\{y_i^k\}$ then follows from Proposition 3.2. □

Note that the approximation criterion (4.8) is implied by the simpler condition

$$\sum_{i=1}^{n}\left(\lambda_i^k\right)^{-2}\|e_i^k\|^2 \le \sigma^2\zeta^2\sum_{i=1}^{n}\|x_i^k - z^k\|^2, \tag{4.25}$$

which might be more convenient to use in practice. The most appropriate way to meet either (4.8) or (4.25) will likely depend on the application. One common situation is that only one of the operators, say, T_1, has a resolvent difficult enough to warrant approximate computation. For example, suppose $F : \mathbb{R}^m \to \mathbb{R}^m$ is a monotone single-valued map, and consider converting the complementarity problem of finding $x \in \mathbb{R}^m$ such that

$$x \ge 0, \qquad F(x) \ge 0, \qquad \langle x, F(x)\rangle = 0 \tag{4.26}$$

to the form $0 \in T_1(x) + T_2(x)$ by setting $T_1 = F$ and $T_2 = N_{\mathbb{R}^m_+}$, as in [10]. Evaluating the resolvent of T_1 then involves solving a possibly large system of linear or nonlinear equations, whereas the resolvent of T_2 is simply projection onto the nonnegative orthant. Thus, we might want to evaluate the resolvent of T_1 approximately using some iterative equation solver, while it is straightforward to evaluate the resolvent of T_2 exactly. A similar situation would occur for the problem

$$\begin{array}{ll} \min & f(x) \\ \text{ST} & x \in C_2 \cap C_3 \cap \cdots \cap C_n, \end{array} \tag{4.27}$$

where $f : \mathbb{R}^m \to \mathbb{R}$ is a twice-differentiable convex function and $C_2,\dots,C_n \subseteq \mathbb{R}^m$ are closed convex sets whose individual projection maps are easy to compute. By setting $T_1 = \nabla f$ and $T_i = N_{C_i}$ for $i = 2,\dots,n$, we obtain a situation where the resolvents of $T_2,\dots,T_n$ are easy to evaluate exactly, but we may want to evaluate the resolvent of T_1 approximately using an iterative unconstrained optimization method. In cases like (4.26) and (4.27), (4.8) simplifies to $\|e_{\pi_k^{-1}(1)}\|^2 \le \sigma^2\kappa(\mathbf{\Lambda}_k^{-1}\mathbf{A}_k)^2\sum_{i=1}^n\|x_i^k - z^k\|^2$.

If more than one operator is a candidate for approximate computation, one simple option would be to require

$$\left(\lambda_i^k\right)^{-2}\left\|e_i^k\right\|^2 \le \sigma^2\kappa\left(\mathbf{\Lambda}_k^{-1}\mathbf{A}_k\right)^2\left\|x_i^k - z^k\right\|^2, \qquad i=1,\ldots,n,$$

since summing these inequalities yields (4.8). However, this approach may be more restrictive than necessary. A less restrictive option would be to interleave iterations for calculating all of the $(x_i,y_i)\in\operatorname{graph}(T_i)$ and terminate as soon as (4.8) itself is satisfied.

5. Variations and special cases.

Rewriting (4.3) as

$$x^k_{\pi_k(i)} + \lambda_i^k y^k_{\pi_k(i)} = z^k + \sum_{j=1}^{i-1}\alpha_{ij}^k\left(x_{\pi_k(j)} - z^k\right) + \lambda_i^k w^k_{\pi_k(i)} + e_i^k,$$

it is natural to consider whether the algorithm could be further generalized by treating the y_i^k in a matter symmetric to the x_i^k. That is, for some β_{ij}^k, $1\le j<i\le n$, one might try to use the $y^k_{\pi_k(j)}$ information generated earlier in the same iteration by replacing (4.3) with

$$x^k_{\pi_k(i)}+\lambda_i^k y^k_{\pi_k(i)} = z^k+\sum_{j=1}^{i-1}\alpha_{ij}^k\left(x_{\pi_k(j)} - z^k\right)+\lambda_i^k\left[w^k_{\pi_k(i)} + \sum_{j=1}^{i-1}\beta_{ij}^k\left(y_{\pi_k(j)} - w^k_{\pi_k(j)}\right)\right]+e_i^k.$$

However, if $e_i^k\equiv 0$, it turns out that this modification does not add any generality to the algorithm. The reason is that it is possible to redefine the α_{ij}^k to obtain an equivalent recursion with $\beta_{ij}^k\equiv 0$. We omit the analysis in the interest of brevity; while more complicated, it resembles that for a similar 2-operator result in [11].

5.1. Including a scaling factor.

A simple variation of the algorithm may be obtained by multiplying the inclusion (1.1) through by any scalar $\eta>0$, arriving at the rescaled formulation $0\in\eta T_1(x)+\cdots+\eta T_n(x)$. Applying Algorithm 3 to this formulation under the substitutions $T_i\leftarrow\eta T_i$, $\lambda_i^k\leftarrow\eta\lambda_i^k$, $y_i^k\leftarrow\eta y_i^k$, and $w_i^k\leftarrow\eta w_i^k$ yields, after some algebraic manipulation, a procedure identical to Algorithm 3, except that (4.10)–(4.12) are modified to incorporate η:

$$y_i^k\in T_i\left(x_i^k\right), \tag{5.1}$$

$$x^k_{\pi_k(i)} + \lambda_i^k y^k_{\pi_k(i)} = \left(1-\sum_{j=1}^{i-1}\alpha_{ij}^k\right)z^k + \sum_{j=1}^{i-1}\alpha_{ij}^k x^k_{\pi_k(j)} + \lambda_i^k w^k_{\pi_k(i)} + e_i^k, \tag{5.2}$$

$$\bar{x}^k = \frac{1}{n}\sum_{i=1}^n x_i^k, \tag{5.3}$$

$$\theta_k = \frac{\sum_{i=1}^n\left\langle z^k - x_i^k, y_i^k - w_i^k\right\rangle}{\eta\left\|\sum_{i=1}^n y_i^k\right\|^2 + \frac{1}{\eta}\sum_{i=1}^n\left\|x_i^k-\bar{x}^k\right\|^2}, \tag{5.4}$$

$$z^{k+1} = z^k - \rho_k\theta_k\eta\sum_{i=1}^n y_i^k, \tag{5.5}$$

$$w_i^{k+1} = w_i^k - \frac{\rho_k\theta_k}{\eta}\left(x_i^k - \bar{x}^k\right). \tag{5.6}$$

This set of recursions produces sequences guaranteed to converge under the same conditions and in the same manner set forth in Proposition 4.2. Essentially, η sets the relative weight the algorithm ascribes to its two main goals: achieving $\sum_{i=1}^{n} y_i^k = 0$ and achieving $x_1^k = \cdots = x_n^k$. In practice, η could be adjusted as the algorithm runs if it appears that these goals are not properly balanced; however, our analysis guarantees convergence only for fixed η.

Suppose $n = 2$, $e_1^k = e_2^k = 0$ for all $k \geq 0$, and π_k is the identity map on $\{1, 2\}$ for all $k \geq 0$. Then, letting $\eta = 1/\sqrt{2}$ causes (5.1)–(5.6) to reduce precisely, after some changes of notation and minor algebraic manipulations, to the two-operator projective splitting algorithm of [11].

5.2. Spingarn's algorithm. In [27], Spingarn presents a partial inverse method for solving the inclusion $\eta_1 T_1(x) + \eta_2 T_2(x) + \cdots + \eta_n T_n(x) \ni 0$. With $\eta_1 = \cdots = \eta_n$, this method reduces, in the notation of this paper, to the following set of recursions to solve (1.1):

$$y_i^k \in T_i\left(x_i^k\right), \qquad i = 1, \ldots, n, \tag{5.7}$$

$$x_i^k + y_i^k = z^k + w_i^k, \qquad i = 1, \ldots, n, \tag{5.8}$$

$$z^{k+1} = \frac{1}{n} \sum_{i=1}^{n} x_i^k, \tag{5.9}$$

$$w_i^{k+1} = y_i^k - \frac{1}{n} \sum_{j=1}^{n} y_j^k, \qquad i = 1, \ldots, n. \tag{5.10}$$

The resolvent evaluations entailed in (5.7)–(5.8) are in fact the same as suggested for the separator calculation at the end of section 2 of this paper and are clearly a special case of our general recursion (4.3). In fact, we now demonstrate that Spingarn's method (5.7)–(5.10) is a special case of the scaled variant (5.1)–(5.6) of our algorithm. Consider (5.1)–(5.6) with $\lambda_i^k = 1$, $\pi_k(i) = i$, $\alpha_{ij}^k = 0$, $e_i^k = 0$, and $\rho_k = 1$ for all $k \geq 0$ and $1 \leq j < i \leq n$. Then the main resolvent relation (5.2) reduces immediately to (5.8). Rearranging (5.8) into $z^k - x_i^k = y_i^k - w_i^k$, we deduce that the numerator of (5.4) is

$$\sum_{i=1}^{n} \left\langle z^k - x_i^k, y_i^k - w_i^k \right\rangle = \sum_{i=1}^{n} \left\| z^k - x_i^k \right\|^2. \tag{5.11}$$

Now, consider the denominator of (5.4). With regard to the first term, we rewrite (5.8) as $y_i^k = z^k - x_i^k + w_i^k$ and then observe that since $\sum_{i=1}^{n} w_i^k = 0$,

$$\sum_{i=1}^{n} y_i^k = \sum_{i=1}^{n} \left(z^k - x_i^k + w_i^k\right) = nz^k - \sum_{i=1}^{n} x_i^k = n\left(z^k - \bar{x}^k\right). \tag{5.12}$$

With regard to the second term in the denominator of (5.4), we calculate

$$\begin{aligned}\sum_{i=1}^{n}\left\|x_i^k-\bar{x}^k\right\|^2 &= \sum_{i=1}^{n}\left\|\left(x_i^k-z^k\right)-\left(\bar{x}^k-z^k\right)\right\|^2\\ &= \sum_{i=1}^{n}\left\|x_i^k-z^k\right\|^2-2\left\langle\sum_{i=1}^{n}\left(x_i^k-z^k\right),\bar{x}^k-z^k\right\rangle+n\left\|\bar{x}^k-z^k\right\|^2\\ &= \sum_{i=1}^{n}\left\|x_i^k-z^k\right\|^2-2\left\langle n\left(\bar{x}^k-z^k\right),\bar{x}^k-z^k\right\rangle+n\left\|\bar{x}^k-z^k\right\|^2\\ &= \sum_{i=1}^{n}\left\|x_i^k-z^k\right\|^2-n\left\|\bar{x}^k-z^k\right\|^2. \qquad (5.13)\end{aligned}$$

Using (5.12) and (5.13), we calculate that the denominator of (5.4) equals

$$\begin{aligned}&n^2\eta\left\|z^k-\bar{x}^k\right\|^2+\frac{1}{\eta}\left(\sum_{i=1}^{n}\left\|x_i^k-z^k\right\|^2-n\left\|\bar{x}^k-z^k\right\|^2\right)\\ &\quad=\left(n^2\eta-\frac{n}{\eta}\right)\left\|z^k-\bar{x}^k\right\|^2+\frac{1}{\eta}\sum_{i=1}^{n}\left\|x_i^k-z^k\right\|^2. \qquad (5.14)\end{aligned}$$

Solving the equation $n^2\eta - n/\eta = 0$, we conclude that the first term in (5.14) will vanish if $\eta = 1/\sqrt{n}$. Combining (5.4), (5.11), and (5.14) with $\eta = 1/\sqrt{n}$, we obtain

$$\theta_k = \frac{\sum_{i=1}^{n}\left\|x_i^k-z^k\right\|^2}{\frac{1}{\eta}\sum_{i=1}^{n}\left\|x_i^k-z^k\right\|^2} = \eta = \frac{1}{\sqrt{n}},$$

unless the denominator is zero, in which case $(z^k, w_1^k, \ldots, w_n^k)$ is already a solution to (1.1). Substituting $\rho_k = 1$, $\theta_k = \eta = 1/\sqrt{n}$, and (5.12) into (5.5), we obtain

$$z^{k+1} = z^k - \frac{1}{n}\sum_{i=1}^{n} y_i^k = z^k - \frac{1}{n}\left(n\left(z^k-\bar{x}^k\right)\right) = \bar{x}^k,$$

which is identical to (5.9). Similarly substituting $\rho_k = 1$ and $\theta_k = \eta = 1/\sqrt{n}$ into (5.6) yields $w_i^{k+1} = w_i^k - (\eta/\eta)(x_i^k - \bar{x}^k) = w_i^k - x_i^k + \bar{x}^k$. From (5.8), $w_i^k - x_i^k = y_i^k - z^k$, so from the definition of $\bar{x}^k$ we then have $w_i^{k+1} = y_i^k - z^k + \frac{1}{n}\sum_{i=1}^{n} x_i^k$. Finally, we rearrange (5.8) into $x_i^k = z^k + w_i^k - y_i^k$ and obtain, using $\sum_{i=1}^{n} w_i^k = 0$, that

$$w_i^{k+1} = y_i^k - z^k + \frac{1}{n}\sum_{j=1}^{n}\left(z^k + w_j^k - y_j^k\right) = y_i^k - z^k + \frac{1}{n}\left(nz^k - \sum_{j=1}^{n} y_j^k\right) = y_i^k - \frac{1}{n}\sum_{j=1}^{n} y_j^k,$$

which is identical to (5.10). Thus, we conclude that, with $\lambda_i^k = 1$, $\pi_k(i) = i$, $\alpha_{ij}^k = 0$, $e_i^k = 0$, $\rho_k = 1$, and $\eta = 1/\sqrt{n}$, the scaled projective algorithm (5.1)–(5.6) reduces exactly to Spingarn's algorithm (5.7)–(5.10).

6. Rudimentary computational experiments.

To demonstrate our algorithmic framework's potential to produce more rapidly converging splitting algorithms, we now describe preliminary experiments on a very simple class of example problems: consider vectors $c^1, \ldots, c^n \in \mathbb{R}^m$, scalars $r_1, \ldots, r_n \geq 0$, and the closed balls

$$B_i \overset{\text{def}}{=} \left\{x \in \mathbb{R}^m \;\middle|\; \|x - c^i\| \leq r_i\right\}, \qquad i = 1, \ldots, n.$$

We consider the nonsmooth convex optimization problem of minimizing the sum of the distances from a point $x \in \mathbb{R}^m$ to all of the balls $B_1, \ldots, B_n$:

$$\min_{x \in \mathbb{R}^m} \sum_{i=1}^{n} \operatorname{dist}(x, B_i), \tag{6.1}$$

where $\operatorname{dist}(x, Y) \overset{\text{def}}{=} \inf\{\|x - y\| \mid y \in Y\}$. We may write such problems in the form (1.1) by letting $T_i = \partial \operatorname{dist}(\cdot, B_i)$, $i = 1, \ldots, n$, where ∂ denotes the subgradient mapping; the resolvent of each such T_i can be evaluated by a simple algorithm (requiring about 10 lines of code in MATLAB, for example). We created artificial problem instances of the form (6.1) by randomly generating the centers c^i uniformly over $[0, 10]^m$ and the radii r_i uniformly over $[0, 2]$.

We solved these problems by both Spingarn's method and by Algorithm 3, with most of Algorithm 3's parameters chosen in a random manner. It is, of course, likely that there are much better ways to choose Algorithm 3's parameters, but random choices give some idea of what may be possible.

To accelerate Spingarn's method, we did not fix ρ_k to 1 as in section 5.2 but experimented with various values in $[0, 2]$; this variation on Spingarn's method does not require the entire projective framework of this paper but may be found in [9] and later works such as [10]. Furthermore, by rescaling the problem $0 \in T_1(x) + \cdots + T_n(x)$ to $0 \in \eta T_1(x) + \cdots + \eta T_n(x)$ as in section 5.1, we generalized (5.8) to

$$x_i^k + \eta y_i^k = z^k + \eta w_i^k, \qquad i = 1, \ldots, n,$$

where $\eta > 0$ is a fixed scalar. These standard, minor generalizations gives Spingarn's method essentially two parameters, $\eta > 0$ and the sequence $\{\rho_k\} \subset [0, 2]$. After a modest amount of experimentation, we found that the values $\eta = 2.5$ and $\rho_k = 1.9$ for all k seemed to yield the fastest results for the problems we generated.

For Algorithm 3, we used a different random permutation π_k of $\{1, \ldots, n\}$ at every iteration. At each iteration, we also independently randomly generated the proximal parameters λ_i^k from the uniform distribution over $[2.5, 3.5]$. To randomly generate the parameters α_{ij}^k, we used the following procedure at each iteration k and for each row $i = 2, \ldots, n$:

- Generate a uniformly distributed random value s_i^k over the interval $[0, .95]$,
- Next, generate $i - 1$ uniform independent random values $\hat{\alpha}_{ij}^k$, $j = 1, \ldots, i - 1$, over the interval $[0, s_i^k]$.
- Finally, obtain the vector $(\alpha_{i1}^k, \ldots, \alpha_{i,i-1}^k)$ by projection of the vector $(\hat{\alpha}_{i1}^k, \ldots, \hat{\alpha}_{i,i-1}^k)$ onto the $(i - 1)$-dimensional simplex

$$\left\{(t_1, \ldots, t_{i-1}) \geq 0 \;\middle|\; t_1 + \cdots + t_{i-1} = s_i^k \right\}.$$

This procedure guarantees that the matrices $\mathbf{A}_k$ and $\mathbf{\Lambda}_k^{-1}\mathbf{A}_k$ are diagonally dominant and hence that $\kappa(\mathbf{\Lambda}_k^{-1}\mathbf{A}_k)$ is bounded away from zero. Clearly, it also yields choices of $\mathbf{\Lambda}_k^{-1}\mathbf{A}_k$ that are bounded, thus meeting the assumptions of Algorithm 3. With the other parameters set in this manner, we experimented with various fixed values of the parameters ρ_k and found that $\rho_k \equiv 1.4$ seemed to work well. We did not experiment with the scaling factor η of (5.1)–(5.6), effectively letting $\eta = 1$.

We applied both algorithms to a single problem instance of each of the sizes $n = m = 100, 200, 300, 400, 500$, terminating when

$$\sqrt{\left\|\sum_{i=1}^{n} y_i^k\right\|^2 + \sum_{i=1}^{n} \left\|x_i^k - \bar{x}^k\right\|^2} \leq 10^{-6}.$$

Table 6.1
Number of iterations for Spingarn's algorithm and Algorithm 3.

n	m	Spingarn	Algorithm 3
100	100	128	31
200	200	176	38
300	300	219	48
400	400	256	56
500	500	289	62

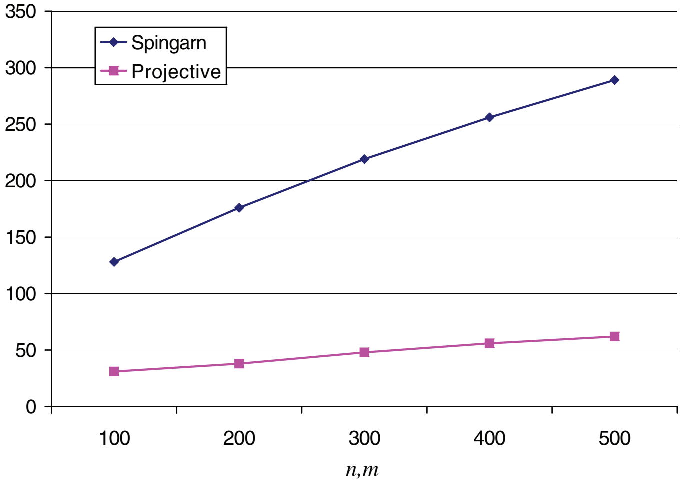


Fig. 6.1. *Number of iterations for Spingarn's algorithm and Algorithm* 3 *("projective").*

The numbers of iterations required by the two algorithms are displayed in Table 6.1 and Figure 6.1; Algorithm 3 appears to require approximately 75–80% fewer iterations for all 5 problems. These results suggest that our projective framework has the potential to significantly accelerate monotone operator splitting methods. To confirm this hypothesis, the next step should be to experiment with a larger sample of problems derived from realistic applications.

7. Conclusions and possible future research. We have proved convergence of a very general class of projective splitting algorithms, extending the results of [11] by allowing for more than two operators, changing the order of operator evaluation with every iteration, and approximate calculation of resolvents using a "relative error" criterion. Some very rudimentary experiments suggest that our framework could significantly improve the convergence of monotone operator splitting applications; however, this hypothesis must be tested in more demanding and realistic settings. Another key issue is to gain insight into how best to take advantage of our framework's new flexibility and larger number of parameters. Rather than setting most of the parameters randomly as in section 6, one could imagine adjusting them to optimize some convergence criterion or error bound which might depend on the application. To this end, some further analysis to attempt to obtain guarantees on the *rate* of convergence would also be of interest. Intuitively, attempting to maximize $\varphi_k(p^k)/\|\nabla\varphi_k\|^2$, and thus the effective constant ξ in Propositions 3.2 and 4.2, would seem to accelerate convergence, but the topic requires further study. Furthermore, the experiments in section 6 do not include the additional scaling parameter η discussed in section 5.1; its properties should be investigated too. Numerical experiments with parameter settings outside the convergence zone established in Proposition 4.2 might also be of interest; for example, it should be possible to weaken the condition $\kappa(\mathbf{\Lambda}_k^{-1}\mathbf{A}_k) > \zeta$ for all $k \geq 0$ to $\limsup_{k\to\infty} \kappa(\mathbf{\Lambda}_k^{-1}\mathbf{A}_k) > 0$.

There also remain many further topics to explore. At the core of our algorithm is a simple projection method using separating hyperplanes. One possible topic for future research is to introduce known variations on this basic projection method into the overall algorithmic setting. For instance, one could examine applying the techniques of [25] to force strong convergence and perhaps to improve practical finite-dimensonal convergence behavior—note that since Corollary 2.3 guarantees that $\operatorname{int}_V S_{\mathrm{e}}(T_1, \dots, T_n) = \emptyset$ in all cases, one cannot entertain techniques like those in [19, 20], which require the solution set to have nonempty interior to obtain strong convergence. In the spirit of projection methods for the intersection of multiple convex sets [2], one might also consider caching some subset of the separating hyperplanes obtained and reprojecting onto them if they are violated in later iterations. In this same vein, one might also consider special, more flexible treatment for those operators that are of the form N_C, where C is a closed convex set, or use of nonorthogonal projections such as those based on Bregman distances; see [5] and numerous later works such as [1]. Finally, a local convergence theory for nonmonotone problems would also be of great interest.

Appendix A. A technical result for infinite dimension.

PROPOSITION A.1. *Let $T_1, \dots, T_n : \mathcal{H} \rightrightarrows \mathcal{H}$ be maximal monotone, and suppose that their sum $T_1 + \cdots + T_n$ is also maximal. Suppose that $z, w_1, \dots, w_n \in \mathcal{H}$ and the sequences $\{(x_i^k, y_i^k)\}_{k=1}^\infty \subset \operatorname{graph}(T_i)$, $i = 1, \dots, n$, have the properties*

$$\left(x_i^k, y_i^k\right) \xrightarrow{\mathsf{w}} (z, w_i), \qquad i = 1, \dots, n, \tag{A.1}$$

$$\sum_{i=1}^n y_i^k \to 0, \tag{A.2}$$

$$\|x_i^k - x_j^k\| \to 0, \qquad i, j = 1, \dots, n. \tag{A.3}$$

Then $(z, w_1, \dots, w_n) \in S_{\mathrm{e}}(T_1, \dots, T_n)$.

Proof. First we claim that

$$0 \in (T_1 + \cdots + T_n)(z). \tag{A.4}$$

To prove this claim, take an arbitrary $(z', w') \in \operatorname{graph}(T_1 + \cdots + T_n)$. Then there exist points $w_i' \in T_i(z')$, $i = 1, \dots, n$, such that $w' = \sum_{i=1}^n w_i'$. Since all of the T_i's are monotone,

$$\left\langle x_i^k - z', y_i^k - w_i' \right\rangle \geq 0, \qquad i = 1, \dots, n. \tag{A.5}$$

Define $\bar{y}^k = \sum_{i=1}^n y_i^k$, and fix any $j \in \{1, \dots, n\}$. We may rewrite the $i = j$ case of (A.5) as

$$\left\langle x_j^k - z', \bar{y}^k - w' + \sum_{\substack{i=1 \\ i \neq j}}^n \left(w_i' - y_i^k\right) \right\rangle \geq 0,$$

and so

$$\left\langle x_j^k - z', -w' \right\rangle \geq -\left\langle x_j^k - z', \bar{y}^k \right\rangle + \sum_{\substack{i=1 \\ i \neq j}}^n \left\langle x_j^k - z', y_i^k - w_i' \right\rangle. \tag{A.6}$$

For any $i \neq j$, we have, courtesy of (A.5), that

$$\left\langle x_j^k - z', y_i^k - w_i' \right\rangle = \left\langle x_i^k - z', y_i^k - w_i' \right\rangle + \left\langle x_j^k - x_i^k, y_i^k - w_i' \right\rangle \geq \left\langle x_j^k - x_i^k, y_i^k - w_i' \right\rangle,$$

and substituting these inequalities into (A.6) yields

$$\left\langle x_j^k - z', -w' \right\rangle \geq -\left\langle x_j^k - z', \bar{y}^k \right\rangle + \sum_{\substack{i=1\\ i\neq j}}^{n} \left\langle x_j^k - x_i^k, y_i^k - w_i' \right\rangle. \tag{A.7}$$

We now consider taking $k \to \infty$ in (A.7). Since $x_j^k \xrightarrow{\text{w}} z$ by (A.1), the limit of the left-hand side of (A.7) is $\langle z - z', -w' \rangle$. Since the weakly convergent sequence $\{x_j^k\}$ must be bounded and $\bar{y}^k \to 0$ by (A.2), we have $\langle x_j^k - z', \bar{y}^k \rangle \to 0$. Similarly, for each $i \neq j$, we have that $x_j^k - x_i^k \to 0$ by (A.3), and the weakly convergent sequence $\{y_i^k\}$ must be bounded so that $\langle x_j^k - x_i^k, y_i^k - w_i' \rangle \to 0$. Thus, taking the limit in (A.7) yields $\langle z - z', -w' \rangle \geq 0$. Since $T_1 + \cdots + T_n$ is maximal monotone and $(z', w') \in \operatorname{graph}(T_1 + \cdots + T_n)$ was arbitrary, we conclude that (A.4) holds.

Next, we claim that

$$\lim_{k\to\infty} \left\langle x_i^k, y_i^k \right\rangle = \langle z, w_i \rangle, \qquad i = 1, \ldots, n. \tag{A.8}$$

In view of (A.4), there must exist $u_i \in T_i(z)$, $i = 1, \ldots, n$, such that $\sum_{i=1}^n u_i = 0$. Since the T_i's are monotone, we have $\langle x_i^k - z, y_i^k - u_i \rangle \geq 0$, $i = 1, \ldots, n$, which we may rearrange to obtain

$$\left\langle x_i^k, y_i^k \right\rangle \geq \left\langle z, y_i^k - u_i \right\rangle + \left\langle x_i^k, u_i \right\rangle, \qquad i = 1, \ldots, n.$$

From (A.1), it is easily deduced that the right-hand sides of the above inequalities converge, respectively, to $\langle z, w_i \rangle$. Hence,

$$\liminf_{k\to\infty} \left\langle x_i^k, y_i^k \right\rangle \geq \langle z, w_i \rangle, \qquad i = 1, \ldots, n. \tag{A.9}$$

Once again, fix some $j \in \{1, \ldots, n\}$. Then we observe that

$$\begin{aligned} \left\langle x_j^k, y_j^k \right\rangle &= \left\langle x_j^k, \bar{y}^k \right\rangle - \sum_{\substack{i=1\\ i\neq j}}^{n} \left\langle x_j^k, y_i^k \right\rangle \\ &= \left\langle x_j^k, \bar{y}^k \right\rangle - \sum_{\substack{i=1\\ i\neq j}}^{n} \left(\left\langle x_i^k, y_i^k \right\rangle + \left\langle x_j^k - x_i^k, y_i^k \right\rangle \right). \end{aligned}$$

We now take the lim sup as $k \to \infty$ of the above equation. Using logic resembling that for (A.7), we observe that $\langle x_j^k, \bar{y}^k \rangle \to 0$ and $\langle x_j^k - x_i^k, y_i^k \rangle \to 0$. Therefore, using (A.9),

$$\limsup_{k\to\infty} \left\langle x_j^k, y_j^k \right\rangle \leq -\sum_{\substack{i=1\\ i\neq j}}^{n} \langle z, w_i \rangle = -\left\langle z, \sum_{\substack{i=1\\ i\neq j}}^{n} w_i \right\rangle. \tag{A.10}$$

Since $y_i^k \xrightarrow{\text{w}} w_i$, $i = 1, \ldots, n$, we have $\sum_{i=1}^n y_i^k \xrightarrow{\text{w}} \sum_{i=1}^n w_i$, and therefore, also using that $\sum_{i=1}^n y_i^k \to 0$,

$$\left\langle \sum_{i=1}^n w_i, \sum_{i=1}^n w_i \right\rangle = \lim_{k\to\infty} \left\langle \sum_{i=1}^n y_i^k, \sum_{i=1}^n w_i \right\rangle = \left\langle \lim_{k\to\infty} \sum_{i=1}^n y_i^k, \sum_{i=1}^n w_i \right\rangle = 0,$$

so we must have $\sum_{i=0}^{n} w_i = 0$. Thus, (A.10) implies $\limsup_{k\to\infty} \langle x_j^k, y_j^k\rangle \leq \langle z, w_j\rangle$, which, combined with (A.9), means that $\lim_{k\to\infty}\langle x_j^k, y_j^k\rangle = \langle z, w_j\rangle$. Since j was arbitrary, (A.8) holds.

Finally, we claim that $w_i \in T_i(z)$, $i = 1, \ldots, n$. To prove this inclusion, take any $i \in \{1, \ldots, n\}$ and $(z', w_i') \in \operatorname{graph}(T_i)$. Then the monotonicity of T_i implies

$$\left\langle x_i^k - z', y_i^k - w_i'\right\rangle = \left\langle x_i^k, y_i^k\right\rangle - \left\langle z', y_i^k\right\rangle - \left\langle x_i^k, w_i'\right\rangle + \langle z', w_i'\rangle \geq 0.$$

Applying (A.1) and (A.8) while taking the limit as $k \to \infty$ yields

$$\langle z, w_i\rangle - \langle z', w_i\rangle - \langle z, w_i'\rangle + \langle z', w_i'\rangle \geq 0\,,$$

which is equivalent to $\langle z - z', w_i - w_i'\rangle \geq 0$. Since the T_i's are maximal and both $(z', w_i') \in \operatorname{graph}(T_i)$ and $i \in \{1, \ldots, n\}$ were arbitrary, we conclude that we indeed have $w_i \in T_i(z)$. Finally, since we have already established that $\sum_{i=0}^{n} w_i = 0$, it follows from $w_i \in T_i(z)$ that we must have $(z, w_1, \ldots, w_n) \in S_{\mathrm{e}}(T_1, \ldots, T_n)$. □

Appendix B. Proof of Lemma 4.1.

Proof. If $u = 0$, then $\mathbf{L}u = 0$ and (4.5)–(4.6) hold trivially, so it remains to consider the case that at least one u_i is nonzero. Given any such u, let $v = (v_1, \ldots, v_n)$ and ℓ_{ij} be defined as in (4.4). Define $U \subseteq \mathcal{H}$ to be the finite-dimensional subspace spanned by $u_1, \ldots, u_n$ in $\mathcal{H}$. From (4.4), we have $v_i \in U$ for $i = 1, \ldots, n$, and, thus, $u, v \in U^n$. Letting $n' \leq n$ denote the dimension of U, take $B = (b_1, \ldots, b_{n'})$ to be some orthonormal basis for U. From B, we may create an orthonormal basis $\overline{B} = (\overline{b}_1, \ldots, \overline{b}_{n'n})$ for U^n via

$$\begin{array}{lll}
\overline{b}_1 = (b_1, 0, 0, \ldots, 0), & \overline{b}_{n+1} = (b_2, 0, 0, \ldots, 0) \quad \cdots & \overline{b}_{(n'-1)n+1} = (b_{n'}, 0, 0, \ldots, 0),\\
\overline{b}_2 = (0, b_1, 0, \ldots, 0), & \overline{b}_{n+2} = (0, b_2, 0, \ldots, 0) \quad \cdots & \overline{b}_{(n'-1)n+2} = (0, b_{n'}, 0, \ldots, 0),\\
\quad\vdots & \qquad\vdots & \qquad\vdots\\
\overline{b}_n = (0, 0, \ldots, 0, b_1), & \overline{b}_{2n} = (0, 0, \ldots, 0, b_2) \quad \cdots & \overline{b}_{n'n} = (0, 0, \ldots, 0, b_{n'}).
\end{array}$$

Let $\overline{\mathbf{u}} \in \mathbb{R}^{n'n}$ be the unique representation of u with respect to this basis, that is, its elements $\overline{u}_m$, $m = 1, \ldots, n'n$, are such that $u = \sum_{m=1}^{n'n} \overline{u}_m \overline{b}_m$. Similarly, let $\overline{\mathbf{v}} \in \mathbb{R}^{n'n}$ be the unique representation of v. By the orthonormality of the basis $\overline{B}$, it follows that $\|u\| = \|\overline{\mathbf{u}}\|$, $\|v\| = \|\overline{\mathbf{v}}\|$, and $\langle u, \mathbf{L}u\rangle = \langle u, v\rangle = \overline{\mathbf{u}}^\top \overline{\mathbf{v}}$. Let us now examine the action of the linear mapping defined by (4.4) on the basis $\overline{B}$, namely,

$$\begin{aligned}
\overline{b}_1 = (b_1, 0, 0, \ldots, 0) &\mapsto (\ell_{11}b_1, \ell_{21}b_1, \ldots, \ell_{n1}b_1) = \ell_{11}\overline{b}_1 + \ell_{21}\overline{b}_2 + \cdots + \ell_{n1}\overline{b}_n,\\
\overline{b}_2 = (0, b_1, 0, \ldots, 0) &\mapsto (\ell_{12}b_1, \ell_{22}b_1, \ldots, \ell_{n2}b_1) = \ell_{12}\overline{b}_1 + \ell_{22}\overline{b}_2 + \cdots + \ell_{n2}\overline{b}_n,\\
&\ \ \vdots\\
\overline{b}_n = (0, 0, \ldots, 0, b_1) &\mapsto (\ell_{1n}b_1, \ell_{2n}b_1, \ldots, \ell_{nn}b_1) = \ell_{1n}\overline{b}_1 + \ell_{2n}\overline{b}_2 + \cdots + \ell_{nn}\overline{b}_n,\\
\overline{b}_{n+1} = (b_2, 0, 0, \ldots, 0) &\mapsto (\ell_{11}b_2, \ell_{21}b_2, \ldots, \ell_{n1}b_2) = \ell_{11}\overline{b}_{n+1} + \ell_{21}\overline{b}_{n+2} + \cdots + \ell_{n1}\overline{b}_{2n},\\
\overline{b}_{n+2} = (0, b_2, 0, \ldots, 0) &\mapsto (\ell_{12}b_2, \ell_{22}b_2, \ldots, \ell_{n2}b_2) = \ell_{12}\overline{b}_{n+1} + \ell_{22}\overline{b}_{n+2} + \cdots + \ell_{n2}\overline{b}_{2n},\\
&\ \ \vdots\\
\overline{b}_{n'n} = (0, 0, \ldots, 0, b_{n'}) &\mapsto (\ell_{1n}b_{n'}, \ell_{2n}b_{n'}, \ldots, \ell_{nn}b_{n'})\\
&\qquad = \ell_{1n}\overline{b}_{(n'-1)n+1} + \ell_{2n}\overline{b}_{(n'-1)n+2} + \cdots + \ell_{nn}\overline{b}_{n'n}.
\end{aligned}$$

Thus, in terms of the basis $\overline{B}$, the action of the linear mapping (4.4) is that of the $n'n \times n'n$ block-diagonal matrix

$$\overline{\mathbf{L}} \overset{\text{def}}{=} \underbrace{\begin{bmatrix} \mathbf{L} & & & \\ & \mathbf{L} & & \\ & & \ddots & \\ & & & \mathbf{L} \end{bmatrix}}_{n' \text{ times}},$$

and we must have $\overline{\mathbf{v}} = \overline{\mathbf{L}}\overline{\mathbf{u}}$. It is easily seen that $\operatorname{sym} \overline{\mathbf{L}}$ has the same eigenvalues as $\operatorname{sym} \mathbf{L}$, so $\kappa(\overline{\mathbf{L}}) = \kappa(\mathbf{L})$. Using standard eigenvalue analysis in $\mathbb{R}^{n'n}$, we therefore have

$$\overline{\mathbf{u}}^\top \overline{\mathbf{L}}\overline{\mathbf{u}} \geq \kappa(\overline{\mathbf{L}}) \left\|\overline{\mathbf{u}}\right\|^2 = \kappa(\mathbf{L}) \left\|\overline{\mathbf{u}}\right\|^2 .$$

Substituting $\|u\| = \|\overline{\mathbf{u}}\|$ and $\langle u, \mathbf{L}u \rangle = \langle u, v \rangle = \overline{\mathbf{u}}^\top \overline{\mathbf{v}} = \overline{\mathbf{u}}^\top \overline{\mathbf{L}}\overline{\mathbf{u}}$ into this relation yields (4.6). To establish (4.5), we observe that

$$\begin{aligned}
\left\|\overline{\mathbf{L}}\right\|^2 &= \max\left\{ \left\|\overline{\mathbf{L}}\overline{\mathbf{x}}\right\|^2 \;\middle|\; \overline{\mathbf{x}} \in \mathbb{R}^{n'n},\ \|\overline{\mathbf{x}}\| = 1 \right\} \\
&= \max\left\{ \sum_{j=1}^{n'} \|\mathbf{L}\mathbf{x}_j\|^2 \;\middle|\; \mathbf{x}_1, \dots, \mathbf{x}_{n'} \in \mathbb{R}^n,\ \sum_{j=1}^{n'} \|\mathbf{x}_j\|^2 = 1 \right\} \\
&= \max\left\{ \sum_{j=1}^{n'} \max\left\{ \|\mathbf{L}\mathbf{x}\|^2 \;\middle|\; \begin{array}{l} \mathbf{x} \in \mathbb{R}^n \\ \|\mathbf{x}\|^2 = \nu_j \end{array} \right\} \;\middle|\; \begin{array}{l} \nu_1, \dots, \nu_{n'} \geq 0 \\ \nu_1 + \cdots + \nu_{n'} = 1 \end{array} \right\} \\
&= \max\left\{ \sum_{j=1}^{n'} \nu_j \|\mathbf{L}\|^2 \;\middle|\; \begin{array}{l} \nu_1, \dots, \nu_{n'} \geq 0 \\ \nu_1 + \cdots + \nu_{n'} = 1 \end{array} \right\} \; = \; \|\mathbf{L}\|^2 .
\end{aligned}$$

Thus, we may substitute $\left\|\overline{\mathbf{L}}\right\| = \|\mathbf{L}\|$ into the inequality $\left\|\overline{\mathbf{L}}\overline{\mathbf{u}}\right\| \leq \left\|\overline{\mathbf{L}}\right\| \|\overline{\mathbf{u}}\|$, along with $\left\|\overline{\mathbf{L}}\overline{\mathbf{u}}\right\| = \|\overline{\mathbf{v}}\| = \|v\| = \|\mathbf{L}u\|$ and $\|u\| = \|\overline{\mathbf{u}}\|$, to obtain (4.5). □

REFERENCES

[1] H. H. Bauschke, J. M. Borwein, and P. L. Combettes, *Bregman monotone optimization algorithms*, SIAM J. Control Optim., 42 (2003), pp. 596–636.

[2] H. H. Bauschke and J. M. Borwein, *On projection algorithms for solving convex feasibility problems*, SIAM Rev., 38 (1996), pp. 367–426.

[3] H. H. Bauschke, P. L. Combettes, and S. G. Kruk, *Extrapolation algorithm for affine-convex feasibility problems*, Numer. Algorithms, 41 (2006), pp. 239–274.

[4] H. H. Bauschke, P. L. Combettes, and S. Reich, *The asymptotic behavior of the composition of two resolvents*, Nonlinear Anal., 60 (2005), pp. 283–301.

[5] L. M. Brègman, *A relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming*, Žj. Vyčhisl. Mat. Mat. Fiz., 7 (1967), pp. 620–631.

[6] G. Cimmino, *Calcolo approssimato per le soluzioni dei sistemi di equazioni lineari.*, Ric. Sci. Progr. Tecn. Econom. Naz., 1 (1938), pp. 326–333.

[7] P. L. Combettes, *Solving monotone inclusions via compositions of nonexpansive averaged operators*, Optimization, 53 (2004), pp. 475–504.

[8] J. Douglas, Jr., and H. H. Rachford, Jr., *On the numerical solution of heat conduction problems in two and three space variables*, Trans. Amer. Math. Soc., 82 (1956), pp. 421–439.

[9] J. Eckstein and D. P. Bertsekas, *On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators*, Math. Program., 55 (1992), pp. 293–318.

[10] J. Eckstein and M. C. Ferris, *Operator-splitting methods for monotone affine variational inequalities, with a parallel application to optimal control*, INFORMS J. Comput., 10 (1998), pp. 218–235.

[11] J. Eckstein and B. F. Svaiter, *A family of projective splitting methods for the sum of two maximal monotone operators*, Math. Program., 111 (2008), pp. 173–199.

[12] D. Gabay, *Applications of the method of multipliers to variational inequalities*, in Augmented Lagrangian Methods: Applications to the Solution of Boundary Value Problems, M. Fortin and R. Glowinski, eds., North–Holland, Amsterdam, 1983, pp. 299–340.

[13] S. Kaczmarz, *Angenäherte Auflösung von Systemen linearer Gleichungen*, Bull. Int. Acad. Polon. Sci. Let. A, 35 (1937), pp. 355–357.

[14] S. Kaczmarz, *Approximate solution of systems of linear equations*, Internat. J. Control, 57 (1993), pp. 1269–1271.

[15] J. Lawrence and J. E. Spingarn, *On fixed points of nonexpansive piecewise isometric mappings*, Proc. London Math. Soc., 55 (1987), pp. 605–624.

[16] P. L. Lions and B. Mercier, *Splitting algorithms for the sum of two nonlinear operators*, SIAM J. Numer. Anal., 16 (1979), pp. 964–979.

[17] P. L. Lions, *Une méthode itérative de résolution d'une inéquation variationnelle*, Israel J. Math., 31 (1978), pp. 204–208.

[18] G. J. Minty, *Monotone (nonlinear) operators in Hilbert space*, Duke Math. J., 29 (1962), pp. 341–346.

[19] O. Nevanlinna and S. Reich, *Strong convergence of contraction semigroups and of iterative methods for accretive operators in Banach spaces*, Israel J. Math., 32 (1979), pp. 44–58.

[20] N. Ottavy, *Strong convergence of projection-like methods in Hilbert spaces*, J. Optim. Theory Appl., 56 (1988), pp. 433–461.

[21] G. B. Passty, *Ergodic convergence to a zero of the sum of monotone operators in Hilbert space*, J. Math. Anal. Appl., 72 (1979), pp. 383–390.

[22] D. W. Peaceman and H. H. Rachford, Jr., *The numerical solution of parabolic and elliptic differential equations*, J. SIAM, 3 (1955), pp. 28–41.

[23] M. V. Solodov and B. F. Svaiter, *A hybrid approximate extragradient-proximal point algorithm using the enlargement of a maximal monotone operator*, Set-Valued Anal., 7 (1999), pp. 323–345.

[24] M. V. Solodov and B. F. Svaiter, *A hybrid projection-proximal point algorithm*, J. Convex Anal., 6 (1999), pp. 59–70.

[25] M. V. Solodov and B. F. Svaiter, *Forcing strong convergence of proximal point iterations in a Hilbert space*, Math. Program., 87 (2000), pp. 189–202.

[26] M. V. Solodov and B. F. Svaiter, *A unified framework for some inexact proximal point algorithms*, Numer. Funct. Anal. Optim., 22 (2001), pp. 1013–1035.

[27] J. E. Spingarn, *Partial inverse of a monotone operator*, Appl. Math. Optim., 10 (1983), pp. 247–265.

[28] P. Tseng, *A modified forward-backward splitting method for maximal monotone mappings*, SIAM J. Control Optim., 38 (2000), pp. 431–446.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 2, pp. 812–838

© 2009 Society for Industrial and Applied Mathematics

A PROBLEM OF GUARDING LINE SEGMENT[*]

WITOLD RZYMOWSKI[†]

Abstract. A line segment on the plane is guarded by n defenders. One invader wants to pass through the line segment, but he has to keep the distance from each defender no less than a given constant. All defenders can move along a straight line only with maximal speed 1. The invader can move on the whole plane with maximal speed greater than 1. Other than this, no kinematic or dynamic constraints are imposed on the defenders' and the invader's motions. The maximal length of the line segment which can be guarded by defenders is established in this paper.

Key words. differential games, guarding territory problems

AMS subject classification. 49N70

DOI. 10.1137/060655857

1. Introduction. This paper deals with the following problem. An invader tries to reach a line segment guarded by n defenders and avoid capture. By capture, we mean that the distance between the invader and a defender (no matter which one) is less than a given constant $\varrho > 0$ after a finite time. The invader wins the game when he can reach the guarded line segment at a moment t^* and can avoid capture indefinitely. The defenders win the game when, at a finite moment t^*, one of them approaches the invader closer than ϱ or the invader does not enter the guarded line segment for all $t \geq 0$. It is assumed that the invader and the defenders know each other's position.

1.1. Contributions. The motion of every defender $\mathbf{D}_k$, $k = 1, 2, \ldots, n$, is restricted to the straight line $\mathbb{R} \times \{0\}$, and the guarded line segment has the form $[0, \Delta] \times \{0\}$, with a $\Delta > 0$. The invader $\mathbf{I}$ can move on the whole plane $\mathbb{R}^2$. Admissible trajectory of each defender $\mathbf{D}_k$ is represented by a function $x_k : [0, \infty) \to \mathbb{R}$ satisfying the Lipschitz condition with constant 1. The symbol $X(a_k)$ will stand for every $\mathbf{D}_k$'s admissible trajectories satisfying the initial condition $x_k(0) = a_k$. Admissible trajectory of the invader is a function $y : [0, \infty) \to \mathbb{R}^2$ satisfying the Lipschitz condition with a fixed constant $\theta > 1$. The symbol $Y(b)$ will stand for every $\mathbf{I}$'s admissible trajectories satisfying the initial condition $y(0) = b$. Invader $\mathbf{I}$ wins the game if there exists a $t^* \geq 0$ such that

$$y(t^*) \in [0, \Delta] \times \{0\}$$

and

$$\min_{k=1,2,\ldots,n} \|y(t) - (x_k(t), 0)\| \geq \varrho$$

for all $t \geq 0$, where $\varrho > 0$ is a fixed capture radius and $\|\cdot\|$ stands for the euclidean norm in $\mathbb{R}^2$. Defenders $\mathbf{D}_1, \ldots, \mathbf{D}_n$ win the game if

$$y(t) \notin [0, \Delta] \times \{0\}$$

[*] Received by the editors March 31, 2006; accepted for publication (in revised form) October 3, 2008; published electronically February 25, 2009.
http://www.siam.org/journals/sicon/48-2/65585.html

[†] Department of Management, Lublin University of Technology, Nadbystrzycka 38 D, 20-618 Lublin, Poland (w.rzymowski@pollub.pl), and Institute of Mathematics and Computer Science, John Paul II Catholic University of Lublin, Konstantynów 1H, 20-708 Lublin, Poland (witrz@kul.lublin.pl).

for all $t \geq 0$, or if there exists a $t^* \geq 0$ such that

$$\min_{k=1,2,\dots,n} \|y(t^*) - (x_k(t^*), 0)\| < \varrho.$$

The problem is to determine the maximal value of Δ such that every line segment $[0, \Delta'] \times \{0\}$, with $\Delta' < \Delta$, can be guarded successfully by defenders. Similarly to how it was done in [1, Chapter II, section 6], the game will be considered from both the invader's and the defenders' points of view.

Invader's point of view. For every $t \geq 0$, invader $\mathbf{I}$ knows his own position $y(t)$, as well as each defender's position $x_k(t)$ and velocity $x'_k(t)$ (when it exists). Let $b = (b_1, b_2) \in \mathbb{R}^2$ and $(a_1, \dots, a_n) \in \mathbb{R}^n$ stand for the initial position of the invader and the defenders, respectively. The invader has an advantage in velocity which allows us to suppose that

$$b_1 = \min_{k=1,2,\dots,n} a_k - \varrho \quad \text{and} \quad b_2 = 0.$$

Roughly speaking, the invader's trajectory $y(t)$ will be generated by a vector field g depending on $y(t)$ and some $x_k(t)$ and $x'_k(t)$ until a moment $\tau = \tau(x_1, \dots, x_n) > 0$ such that $y_2(\tau) = 0$. Next, he will move with maximal speed along a straight line depending on positions of all players at time τ. Strategies of this kind are a special case of the so-called nonanticipating functions introduced (without name) in [2]. Recall that a function

$$\sigma_{\mathbf{I}} : X(a_1) \times \dots \times X_n(a_n) \to Y(b)$$

is nonanticipating if, for each $t \geq 0$ and all,

$$(x_1, \dots, x_n), (\widehat{x}_1, \dots, \widehat{x}_n) \in X(a_1) \times \dots \times X_n(a_n),$$

with

$$x_k(s) = \widehat{x}_k(s), \ 0 \leq s \leq t, \ k = 1, 2, \dots, n;$$

we have

$$\sigma_{\mathbf{I}}(x_1, \dots, x_n)(s) = \sigma_{\mathbf{I}}(\widehat{x}_1, \dots, \widehat{x}_n)(s), \ 0 \leq s \leq t.$$

It will be shown that the invader wins the game for a $\Delta = \Delta(\varrho)$, where $\Delta(\varrho)$ is defined in section 2. Clearly, he then wins the game for any $\Delta' > \Delta(\varrho)$.

Defenders' point of view. Let us fix an arbitrary $r \in (0, \varrho)$. It is reasonable to suppose now that

$$\|b - (a_k, 0)\| > r, \ k = 1, 2, \dots, n.$$

Moreover, if the point b is placed near a guarded line segment, we may assume that the defenders' initial positions are chosen in a special way described in section 4. A defender's trajectory $x_k(t)$ will be defined by the formula

$$x_k(t) = f_k(y(t))$$

until, for a $\tau_1 = \tau_1(y)$, one of the two following possibilities takes place:

$$\|y(\tau_1) - (x_k(\tau_1), 0)\| = r$$

or

$$y_2(\tau_1) = 0 \text{ and } x_{j-1}(\tau_1) + r < y_2(\tau_1) < x_j(\tau_1) - r$$

for an $j \in \{2, 3, \dots, n\}$. Next, in the case of $\tau_1 < \infty$, all defenders will stay at constant positions $x_k(\tau_1)$, $k = 1, 2, \dots, n$, except in the second case, where players $\mathbf{D}_{j-1}$ and $\mathbf{D}_j$ will apply trajectories

$$x_{j-1}(t) = x_{j-1}(\tau_1) + t - \tau_1 \text{ and } x_j(t) = x_{j-1}(\tau_1) - t + \tau_1,$$

respectively. In this connection, each defender $\mathbf{D}_k$ will apply a nonanticipating function

$$\sigma_{\mathbf{D}_k} : Y(b) \to X(a_k)$$

as his guarding strategy. It will be shown that, for $\Delta = \Delta(r)$, defenders win the game. Thus, by the equality

$$\lim_{r \uparrow \varrho} \Delta(r) = \Delta(\varrho),$$

the number $\Delta(\varrho)$ is the solution of the problem.

1.2. Related works. Most of the work concerning the problem of guarding territory study problems involving the visibility of geometrical shapes, the visibility of a moving object, or capturing an evader in an environment; see, e.g., [3], [4], [5], [6], [7]. Our problem is rather related to old problems considered in [8]; see, e.g., [8, Chapter 9, Examples 9.6.3 and 9.6.4]. Moreover, it should be pointed out that vaguely reminiscent problems are discussed now as a part of the RoboFlag game; see [6], for example. The main tools used in solving pursuit-evasion games are the Hamilton–Jacobi–Isaacs equation or the method of resolving functions; see [8] and [9], respectively. Here we apply, together with standard methods of convex analysis and ordinary differential equations theory, a geometric intuitive method which is useful sometimes when other methods fail to work. A problem of guarding a region with maximal area (by one defender) was solved in [10] with the aid of similar methods.

1.3. Organization. The main result (Theorem 1) and the notation used in the paper are given in section 2. In section 3 a discontinuous vector field g generating invasion strategy $\sigma_{\mathbf{I}}$ is constructed. It is shown there (Proposition 1) that the invader wins the game if $\Delta = \Delta(\varrho)$. For arbitrary $r \in (0, \varrho)$ a guarding strategy $\sigma_{\mathbf{D}}$ is described in section 4. It is shown there (Proposition 2) that the defenders win the game if $\Delta = \Delta(r)$. The short section 5 contains a few remarks on the result obtained and possible directions of research.

2. Notation and main result. Throughout this paper we will use the notation

$$\begin{aligned} h_r &= r \sin\beta_0 = \frac{r}{\sqrt{\theta^2+1}}, \ \delta_r = \frac{r}{\cos\beta_0} = \frac{r\sqrt{\theta^2+1}}{\theta}, \\ \varkappa_r &= h_r(\tan\gamma_0 + \cot\gamma_0), \\ w^- &= (\sin\beta_0, -\cos\beta_0), \ w^+ = (\sin\beta_0, \cos\beta_0), \end{aligned}$$

where $\theta > 1$ is fixed, $\beta_0, \gamma_0 \in \left(0, \frac{\pi}{2}\right)$ are such that

$$\begin{aligned} \cos\beta_0 &= \frac{\theta}{\sqrt{\theta^2+1}}, \ \sin\beta_0 = \frac{1}{\sqrt{\theta^2+1}}, \\ \cos\gamma_0 &= \frac{2\theta}{\theta^2+1}, \ \sin\gamma_0 = \frac{\theta^2-1}{\theta^2+1}, \end{aligned}$$

and $r > 0$ is an arbitrary number. Moreover, for all $r > 0$ and $\beta \geq 0$, we put

$$\begin{aligned}
S_r(\beta) &= \frac{r}{\theta^2 - 1} \int_0^\beta \left(\sqrt{\theta^2 - \cos^2 \alpha} + \sin \alpha \right) d\alpha, \\
s_1(r) &= S_r(\pi - \beta_0), \\
s_2(r) &= S_r(\pi - \beta_0) - S_r(\beta_0), \\
s_3(r) &= \frac{h_r}{\theta \sin \gamma_0} = \frac{r\sqrt{\theta^2 + 1}}{\theta(\theta^2 - 1)}, \\
\Lambda_1(r) &= s_1(r) + r(1 + \cos \beta_0) + s_3(r)\,\theta \cos \gamma_0, \\
\Lambda_2(r) &= s_2(r) + 2s_3(r)\,\theta \cos \gamma_0 + 2r \cos \beta_0,
\end{aligned}$$

and

$$\Delta(r) = \begin{cases} S_r(\pi) + 2r, & \text{when} \quad n = 1, \\ 2\Lambda_1(r) + (n-2)\Lambda_2(r), & \text{when} \quad n \geq 2. \end{cases}$$

It is easy to see that $\Delta : (0, \infty) \to (0, \infty)$ is a strictly increasing continuous function satisfying the conditions

$$\lim_{r \downarrow 0} \Delta(r) = 0 \quad \text{and} \quad \lim_{r \uparrow \infty} \Delta(r) = \infty.$$

We are now able to formulate the main result. For this purpose let us denote by $X(a)$ the set of all functions $x : [0, \infty) \to \mathbb{R}$ satisfying the Lipschitz condition with constant 1 and the initial condition $x(0) = a$. For $a_1, \ldots, a_n \in \mathbb{R}$ we will write

$$X(a_1, \ldots, a_n) = X(a_1) \times \cdots \times X(a_n)$$

and say that the number a_k is an initial position of the defender $\mathbf{D}_k$ although, in fact, the pair $(a_k, 0)$ is such a position. Similarly, denote by $Y(b)$ the set of all $y : [0, \infty) \to \mathbb{R}^2$ satisfying the Lipschitz condition with constant $\theta > 1$ and the initial condition $y(0) = b$.

If initial positions $(a_1, 0), \ldots, (a_n, 0), b \in \mathbb{R}^2$, are fixed, we call any nonanticipating mapping

$$\sigma_{\mathbf{D}} : Y(b) \to X(a_1, \ldots, a_n)$$

a guarding strategy. Analogously, we call any nonanticipating mapping

$$\sigma_{\mathbf{I}} : X(a_1, \ldots, a_n) \to Y(b)$$

an invasion strategy.

Let us fix an $r > 0$. A strategy $\sigma_{\mathbf{D}}$ is said to guard the line segment $[0, \Delta(r)] \times \{0\}$ successfully if, for any $y \in Y(b)$, the relation $y(t) \in [0, \Delta(r)] \times \{0\}$ implies the existence of an $s \geq 0$ such that

$$\min_{i=1,2,\ldots,n} \|y(s) - (x_i(s), 0)\| < \varrho,$$

where $(x_1, \ldots, x_n) = \sigma_{\mathbf{D}}(y)$. Further, we say that a strategy $\sigma_{\mathbf{I}}$ rushes for the line segment $[0, \Delta(r)] \times \{0\}$ successfully if for any $(x_1, \ldots, x_n) \in X(a_1, \ldots, a_n)$ we have

$$\min_{i=1,2,\ldots,n} \|\sigma_{\mathbf{I}}(x_1, \ldots, x_n)(t) - (x_i(t), 0)\| \geq \varrho, \; t \geq 0,$$

and there exists an $s \geq 0$ such that

$$\sigma_{\mathbf{I}}(x_1,\ldots,x_n)(s) \in (0,\Delta(r)) \times \{0\}.$$

Definition 1. *The invader wins the game, with $\Delta = \Delta(r)$, if for all $b \in \mathbb{R}^2$ and $(a_1,\ldots,a_n) \in \mathbb{R}^n$ satisfying the condition*

$$\min_{k=1,2,\ldots,n} \|b-(a_k,0)\| \geq \varrho$$

there exists an invasion strategy $\sigma_{\mathbf{I}}$ rushing for the line segment $[0,\Delta(r)] \times \{0\}$ successfully.

Definition 2. *The defenders win the game, with $\Delta = \Delta(r)$, if for each $b \in \mathbb{R}^2$ there exist $(a_1,\ldots,a_n) \in \mathbb{R}^n$ and a guarding strategy $\sigma_{\mathbf{D}}$ such that*

$$0 \leq a_k \leq \Delta(r),\ k=1,2,\ldots,n,$$

and $\sigma_{\mathbf{D}}$ guards the line segment $[0,\Delta(r)] \times \{0\}$ successfully.

Theorem 1. *Given a capture radius $\varrho > 0$, the following statements hold true:*

1. *The invader wins the game, with $\Delta = \Delta(\varrho)$.*
2. *The defenders win the game, with $\Delta = \Delta(r)$, $r \in (0,\varrho)$.*

The proof of Theorem 1 will be given in sections 3 and 4. It will use the following basic facts from convex analysis.

For all $a=(a_1,a_2)$ and $b=(b_1,b_2) \in \mathbb{R}^2$ and each $\alpha \in \mathbb{R}$, we define

$$\langle a,b\rangle = a_1b_1+a_2b_2,\ \|a\| = \sqrt{\langle a,a\rangle} = \sqrt{a_1^2+a_2^2},$$
$$B(a,r) = \left\{x \in \mathbb{R}^2 :\ \|x-a\| < r\right\},$$
$$Ra = (a_2,-a_1),\ e(\alpha) = (-\cos\alpha, \sin\alpha).$$

The symbols int D, bd D, and conv D will stand for the interior, the boundary, and the convex hull of the set $D \subset \mathbb{R}^2$, respectively. For a nonempty set $Z \subset \mathbb{R}^2$, define

$$\operatorname{dist}(y,Z) = \inf_{z\in Z} \|y-z\|,\ y \in \mathbb{R}^2.$$

It is known that

$$|\operatorname{dist}(y,Z) - \operatorname{dist}(\widetilde{y},Z)| \leq \|y-\widetilde{y}\| \quad \text{for all } y,\widetilde{y} \in \mathbb{R}^2.$$

If $Z \subset \mathbb{R}^2$ is a compact convex set, then, for each $y \in \mathbb{R}^2 \diagdown$ int Z, there exists exactly one $\Pi_Z(y) \in$ bd Z such that

$$\|y-\Pi_Z(y)\| = \operatorname{dist}(y,Z).$$

Moreover, Π_Z is a nonexpansive mapping, i.e.,

$$\|\Pi_Z(y) - \Pi_Z(\widetilde{y})\| \leq \|y-\widetilde{y}\| \quad \text{for all } y,\widetilde{y} \in \mathbb{R}^2.$$

Let $D \subset \mathbb{R}^2$ be a convex compact set with $0 \in$ int D. For each $z \in \mathbb{R}^2 \diagdown \{0\}$ there exist a number $\lambda > 0$ and a vector

$$w(D,z) = (w_1(D,z), w_2(D,z)) \in \mathbb{R}^2$$

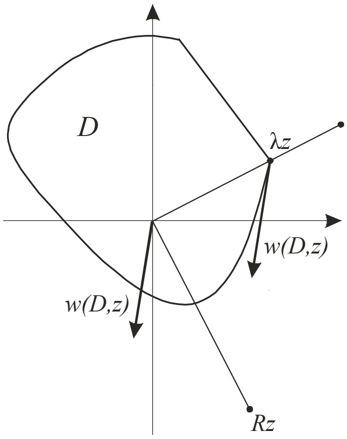


FIG. 1. *Tangent vector $w(D, z)$.*

such that

$$\|w(D, z)\| = 1, \ \langle w(D, z), Rz\rangle > 0$$

and $w(D, z)$ is tangent to the boundary of the set D at the point λz (see Figure 1).

Finally, denote by U the set of all Lebesgue measurable functions

$$u : [0, \infty) \to [-1, 1].$$

We will say that every $u \in U$ is a control. It follows from the well-known Rademacher's theorem that each defender's trajectory x_k is differentiable almost everywhere in $[0, \infty)$ and satisfies the condition

$$|x_k'(t)| \leq 1$$

for almost all $t \in [0, \infty)$. Similarly, in the case of an invader's trajectory y we have

$$\|y'(t)\| \leq \theta$$

for almost all $t \in [0, \infty)$.

3. Attack. For every $k = 1, 2, \ldots, n$, the capture region

$$B((x_k, 0), \varrho) = (x_k, 0) + B(0, \varrho)$$

corresponds to the defender D_k, located at the point $(x_k, 0)$. Clearly, it is an open disc centered at the point $(x_k, 0)$. However, looking at the game from the invader's point of view, we will assign, in fact, to each defender a new extended capture region. It is interesting that the invader should rather keep away from these new artificial regions. It usually occurs that an accurate cooperation gives an individual benefit to each cooperator.

For every $r > 0$ and each $x \in \mathbb{R}$, we define (see Figure 2)

$$D_L(r) = \operatorname{conv}(B[0, r] \cup \{(\delta_r, 0)\}), \ D_L(x, r) = (x, 0) + D_L(r),$$
$$D_M(r) = \operatorname{conv}(B[0, r] \cup \{(\delta_r, 0), (-\delta_r, 0)\}), \ D_M(x, r) = (x, 0) + D_M(r),$$
$$D_R(r) = \operatorname{conv}(B[0, r] \cup \{(-\delta_r, 0)\}), \ D_R(x, r) = (x, 0) + D_R(r),$$

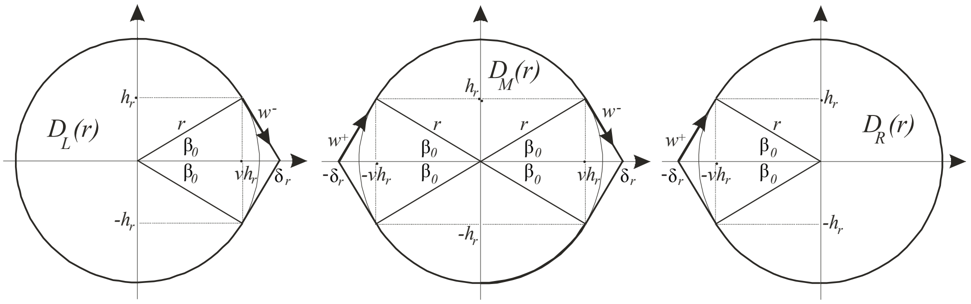


FIG. 2. *Extended capture regions.*

where

$$\xi + Z = \{\xi + z : \ z \in Z\}$$

for a nonempty set $Z \subset \mathbb{R}^2$ and a $\xi \in \mathbb{R}^2$.

Using sets $D_L(r)$, $D_M(r)$, and $D_R(r)$ we will construct an invasion strategy for the invader and guarding strategies for the defenders. The invasion strategy mentioned above will be constructed with the aid of a discontinuous vector field. For this reason we need a few lemmas.

LEMMA 1. *Let us define, for each $z \in \mathbb{R}^2 \setminus \{0\}$ and for every $u \in [-1,1]$,*

$$\psi_L(z,u) = \left(\sqrt{\theta^2 - u^2 w_2^2(D_L(r), z)} - u w_1(D_L(r), z)\right) w(D_L(r), z),$$

and let us fix a control $u \in U$.

1. *The Cauchy problem*

$$\alpha' = \frac{1}{r}\left(\sqrt{\theta^2 - u^2\cos^2\alpha} - u\sin\alpha\right), \ \alpha(0) = 0, \tag{3.1}$$

has exactly one solution $\alpha(u,\cdot) : [0,\infty) \to [0,\infty)$, and there exists a $\tau_{L1}(u)$ such that

$$\tau_{L1}(u) = \min\{t > 0 : \ \alpha(u,t) = \pi - \beta_0\} \le s_1(r).$$

2. *Let us define, for all $t \ge 0$,*

$$v(u,t) = \sqrt{\theta^2 - u^2(t)\cos^2\beta_0} - u(t)\sin\beta_0.$$

There exists exactly one $\tau_{L2}(u) > \tau_{L1}(u)$ such that

$$\int_{\tau_{L1}(u)}^{\tau_{L2}(u)} v(u,s)\,ds = \sqrt{\delta_r^2 - r^2} \tag{3.2}$$

and

$$\tau_{L2}(u) - \tau_{L1}(u) \le s_3(r).$$

3. *The function $z_L(u,\cdot) : [0,\tau_{L2}(u)] \to \mathbb{R}^2$, given by the formulae*

$$z_L(u,t) = re(\alpha(u,t)) \quad \textit{for } t \in [0,\tau_{L1}(u)],$$

$$z_L(u,t) = re(\pi - \beta_0) + \left(\int_{\tau_{L1}(u)}^{t} v(u,s)\,ds\right) w^- \quad \textit{for } t \in [\tau_{L1}(u), \tau_{L2}(u)],$$

is a unique solution to the Cauchy problem

$$z' = \psi_L(z,u), \ z(0) = re(0), \tag{3.3}$$

and satisfies the conditions

$$z_L(u,t) \in \text{bd } D_L(r), \ t \in [0, \tau_{L2}(u)], \quad z_L(u,\tau_{L2}(u)) = \delta_r e(\pi).$$

Proof. Let us fix a control $u \in U$. Note that

$$\begin{aligned}\sqrt{\theta^2 - u^2\cos^2\alpha} - u\sin\alpha &= \frac{\theta^2 - u^2}{\sqrt{\theta^2 - u^2\cos^2\alpha} + u\sin\alpha}\\ &\geq \frac{\theta^2-1}{\theta+1} = \theta - 1 > 0\end{aligned}$$

for all $(u,\alpha) \in [-1,1] \times \mathbb{R}$ and the right-hand side of the differential equation (3.1) is Lipschitzian in α. Thus, there exist exactly one solution $\alpha(u,\cdot) : [0,\infty) \to [0,\infty)$ of (3.1) and a number $\tau_{L1}(u) > 0$ such that

$$\tau_{L1}(u) = \min\{t > 0 : \ \alpha(u,t) = \pi - \beta_0\}.$$

Moreover, we have

$$\begin{aligned}\tau_{L1}(u) = \int_0^{\tau_{L1}(u)} dt &= r\int_0^{\tau_{L1}(u)} \frac{\alpha'(u,t)\,dt}{\sqrt{\theta^2 - u^2(t)\cos^2\alpha(u,t)} - u(t)\sin\alpha(u,t)}\\ &= r\int_0^{\tau_{L1}(u)} \frac{\sqrt{\theta^2 - u^2(t)\cos^2\alpha(u,t)} + u(t)\sin\alpha(u,t)}{\theta^2 - u^2(t)}\alpha'(u,t)\,dt\\ &\leq \frac{r}{\theta^2-1}\int_0^{\tau_{L1}(u)} \left(\sqrt{\theta^2 - \cos^2\alpha(u,t)} + \sin\alpha(u,t)\right)\alpha'(u,t)\,dt\\ &= \frac{r}{\theta^2-1}\int_0^{\pi-\beta_0} \left(\sqrt{\theta^2-\cos^2\beta} + \sin\beta\right)d\beta\\ &= s_1(r),\end{aligned}$$

which finishes the proof of part 1.

Obviously, there exists exactly one $\tau_{L2}(u) > \tau_{L1}(u)$ satisfying condition (3.2). Since

$$\begin{aligned}\frac{r}{\theta} = \sqrt{\delta_r^2 - r^2} &= \int_{\tau_{L1}(u)}^{\tau_{L2}(u)} v(u,s)\,ds\\ &\geq \int_{\tau_{L1}(u)}^{\tau_{L2}(u)} \left(\sqrt{\theta^2 - \cos^2\beta_0} - \sin\beta_0\right)ds\\ &= \frac{\theta^2-1}{\sqrt{\theta^2+1}}\left(\tau_{L2}(u) - \tau_{L1}(u)\right)\end{aligned}$$

we obtain

$$\tau_{L2}(u) - \tau_{L1}(u) \leq \frac{r}{\theta}\Big/\frac{\theta^2-1}{\sqrt{\theta^2+1}} = s_3(r),$$

which finishes the proof of part 2. □

In order to establish the local existence and uniqueness of the Cauchy problem (3.3), with arbitrary initial condition $z(0) \neq 0$, it is enough to see that the vector field ψ_L is locally Lipschitzian with respect to the variable z in the set $\mathbb{R}^2 \diagdown ([0,\infty) \times \{0\})$ and transversal to the ray $(0,\infty) \times \{0\}$. It follows from the relations

$$w(D_L(r), re(\alpha(t))) = Re(\alpha(t)) = (\sin\alpha(t), \cos\alpha(t)), \ t \in [0, \tau_{L1}(u)],$$

that the function $z_L(u,\cdot)$ solves problem (3.3) in the interval $[0,\tau_{L1}(u)]$. Analogously, $z_L(u,\cdot)$ solves problem (3.3) in the interval $[\tau_{L1}(u), \tau_{L2}(u)]$ since

$$w(D_L(r), z_L(u,t)) = w^- = (\sin\beta_0, -\cos\beta_0), \ t \in [\tau_{L1}(u), \tau_{L2}(u)).$$

The remaining properties of $z_L(u,\cdot)$ are easy to prove.

The next two lemmas, with $v(u,\cdot)$ defined in Lemma 1, have proofs analogous to the proof of Lemma 1.

Lemma 2. *Let us define, for each $z \in \mathbb{R}^2 \diagdown \{0\}$ and for every $u \in [-1,1]$,*

$$\psi_M(z,u) = \left(\sqrt{\theta^2 - u^2 w_2^2(D_M(r), z)} - uw_1(D_M(r), z)\right) w(D_M(r), z),$$

and let us fix a control $u \in U$. Then the following statements hold true:

1. *There exists exactly one $\tau_{M1}(u) > 0$ such that*

$$\int_0^{\tau_{M1}(u)} v(u,s)\,ds = \sqrt{\delta_r^2 - r^2} \ \text{ and } \ \tau_{M1}(u) \leq s_3(r).$$

2. *The Cauchy problem*

$$\alpha' = \frac{1}{r}\left(\sqrt{\theta^2 - u^2\cos^2\alpha} - u\sin\alpha\right), \ \alpha(\tau_{M1}(u)) = \beta_0,$$

has exactly one solution $\alpha(u,\cdot) : [\tau_{M1}, \infty) \to [\beta_0, \infty)$, and there exists a $\tau_{M2}(u)$ such that

$$\tau_{M2}(u) = \min\{t > \tau_{M1} : \ \alpha(u,t) = \pi - \beta_0\} \leq \tau_{M1}(u) + s_2(r).$$

3. *There exists exactly one $\tau_{M3}(u) > \tau_{M2}(u)$ such that*

$$\int_{\tau_{M2}(u)}^{\tau_{M3}(u)} v(u,s)\,ds = \sqrt{\delta_r^2 - r^2} \ \text{ and } \ \tau_{M3}(u) \leq \tau_{M2}(u) + s_3(r).$$

4. *The function $z_M(u,\cdot) : [0, \tau_{M3}(u)] \to \mathbb{R}^2$, given by the formulae*

$$z_M(u,t) = \delta_r e(0) + \left(\int_0^t v(u,s)\,ds\right) w^+, \ t \in [0, \tau_{M1}(u)],$$
$$z_M(u,t) = re(\alpha(u,t)) \quad \text{for } t \in [\tau_{M1}(u), \tau_{M2}(u)],$$
$$z_M(u,t) = re(\pi - \beta_0) + \left(\int_{\tau_{M2}(u)}^t v(u,s)\,ds\right) w^-, \ t \in [\tau_{M2}(u), \tau_{M3}(u)],$$

is a unique solution to the Cauchy problem

$$z' = \psi_M(z,u), \ z(0) = \delta_r e(0),$$

and satisfies the conditions

$$z_M(u,t) \in \text{bd } D_M(r), \ t \in [0, \tau_{M3}(u)], \quad z_M(u, \tau_{M3}(u)) = \delta_r e(\pi).$$

LEMMA 3. *Let us define, for each $z \in \mathbb{R}^2 \diagdown \{0\}$ and for every $u \in [-1,1]$,*

$$\psi_R(z,u) = \left(\sqrt{\theta^2 - u^2 w_2^2(D_R(r),z)} - uw_1(D_R(r),z)\right) w(D_R(r),z),$$

and let us fix a control $u \in U$. Then the following statements hold true:

1. *There exists exactly one $\tau_{R1}(u) > 0$ such that*

$$\int_0^{\tau_{R1}(u)} v(u,s)\,ds = \sqrt{\delta_r^2 - r^2} \quad \text{and} \quad \tau_{R1}(u) \le s_3(r).$$

2. *The Cauchy problem*

$$\alpha' = \frac{1}{r}\left(\sqrt{\theta^2 - u^2\cos^2\alpha} - u\sin\alpha\right), \ \alpha(\tau_{R1}(u)) = \beta_0,$$

has exactly one solution $\alpha(u,\cdot) : [\tau_{R1}(u),\infty) \to [\beta_0,\infty)$, and there exists a $\tau_{R2}(u)$ such that

$$\tau_{R2}(u) = \min\{t > \tau_{R1} : \ \alpha(u,t) = \pi\} \le \tau_{R1}(u) + s_1(r).$$

3. *The function $z_R(u,\cdot) : [0,\tau_{R2}(u)] \to \mathbb{R}^2$, given by the formula*

$$z_R(u,t) = \delta_r e(0) + \left(\int_0^t v(u,s)\,ds\right) w^+, \ t \in [0,\tau_{R1}(u)],$$
$$z_R(u,t) = re(\alpha(u,t)) \quad \text{for } t \in [\tau_{R1}(u),\tau_{R2}(u)],$$

is a unique solution to the Cauchy problem

$$z' = \psi_R(z,u), \ z(0) = \delta_r e(0),$$

and satisfies the conditions

$$z_R(u,t) \in \operatorname{bd} D_R(r), \ t \in [0,\tau_{R2}(u)], \quad z_R(u,\tau_{R2}(u)) = re(\pi).$$

Now we shall formulate a corollary playing a crucial role in what follows. Let D be one of the sets $D_L(r)$, $D_M(r)$, $D_R(r)$. For all $y \in \mathbb{R}^2$ and $x \in \mathbb{R}$, with $y \neq (x,0)$, and for all $u \in [-1,1]$, we set

$$\begin{aligned} g_D(y,x,u) &= uw_2(D,y-(x,0))\,Rw(D,y-(x,0)) \\ &\quad + \sqrt{\theta^2 - u^2 w_2^2(D,y-(x,0))}\,w(D,y-(x,0)). \end{aligned}$$

Note that

$$uw_2(D,y-(x,0)) = \langle (u,0), Rw(D,y-(x,0))\rangle$$

and

$$\begin{aligned} (u,0) &= \langle (u,0), Rw(D,y-(x,0))\rangle\, Rw(D,y-(x,0)) \\ &\quad + \langle (u,0), w(D,y-(x,0))\rangle\, w(D,y-(x,0)) \\ &= uw_2(D,y-(x,0))\,Rw(D,y-(x,0)) \\ &\quad - uw_1(D,z)\,w(D,y-(x,0)) \end{aligned}$$

for each $u \in [-1,1]$. Employing this observation and applying Lemma 1, one can check easily that, for each $a \in \mathbb{R}$ and each $x \in X(a)$, the function

$$y(t) = (x(t), 0) + z_L(x', t), \ t \geq 0,$$

is a unique solution to the Cauchy problem

$$y' = g_{D_L(r)}(y, x, x'), \ y(0) = (a, 0) + re(0),$$

in the whole interval $[0, \infty)$. Moreover it satisfies the following conditions:

$$y(\tau_{L1}(x')) = (x(\tau_{L1}(x')), 0) + re(\pi - \beta_0), \ y(\tau_{L2}(x')) = (x(\tau_{L2}(x')), 0) + \delta_r e(\pi),$$

and

$$y(t) - (x(t), 0) \in \text{bd } D_L(r), \ t \geq 0.$$

Since the analogous statements hold true with $g_{D_L(r)}$ replaced by $g_{D_M(r)}$ and $g_{D_R(r)}$, we obtain the following.

Corollary 1. *Let us fix an $a \in \mathbb{R}$ and an $x \in X(a)$. Then the following statements hold true:*

1. *The function*

$$y_L(x,t) = (y_{L1}(x,t), y_{L2}(x,t)) = (x(t), 0) + z_L(x', t), \ t \geq 0,$$

is a unique solution to the Cauchy problem

$$y' = g_{D_L(r)}(y, x, x'), \ y(0) = (a, 0) + re(0),$$

in the whole interval $[0, \infty)$, and satisfies the following conditions:

$$y_L(x,t) - (x(t), 0) \in \text{bd } D_L(r), \ t \geq 0,$$

$$\begin{aligned} y_L(x, \tau_{L1}(x')) = (x(\tau_{L1}(x')), 0) + re(\pi - \beta_0), \ & y_L(x, \tau_{L2}(x')) \\ = (x(\tau_{L2}(x')), 0) + \delta_r e(\pi), & \end{aligned}$$

$$\begin{aligned} y_{L1}(x, \tau_{L1}(x')) - y_{L1}(x, 0) &\leq s_1(r) + r(1 + \cos\beta_0), \\ y_{L1}(x, \tau_{L2}(x')) - y_{L1}(x, \tau_{L1}(x')) &\leq s_3(r)\,\theta\cos\gamma_0. \end{aligned}$$

2. *The function*

$$y_M(x,t) = (y_{M1}(x,t), y_{M2}(x,t)) = (x(t), 0) + z_M(x', t), \ t \geq 0,$$

is a unique solution to the Cauchy problem

$$y' = g_{D_M(r)}(y, x, x'), \ y(0) = (a, 0) + \delta_r e(0),$$

in the whole interval $[0, \infty)$, and satisfies the following conditions:

$$y(x,t) - (x(t), 0) \in \text{bd } D_L(r), \ t \geq 0,$$

$$\begin{aligned} y_M\left(x,\tau_{M1}\left(x'\right)\right) &= \left(x\left(\tau_{M1}\left(x'\right)\right),0\right)+re\left(\beta_0\right),\ y_M\left(x,\tau_{M2}\left(x'\right)\right)\\ &= \left(x\left(\tau_{M2}\left(x'\right)\right),0\right)+re\left(\pi-\beta_0\right),\\ y_M\left(x,\tau_{M3}\left(x'\right)\right) &= \left(x\left(\tau_{M3}\left(x'\right)\right),0\right)+\delta_r e\left(\pi\right),\end{aligned}$$

$$\begin{aligned} y_{M1}\left(x,\tau_{M1}\left(x'\right)\right)-y_{M1}\left(x,0\right) &\le s_3\left(r\right)\theta\cos\gamma_0,\\ y_{M1}\left(x,\tau_{M2}\left(x'\right)\right)-y_{M1}\left(x,\tau_{M1}\left(x'\right)\right) &\le s_2\left(r\right)+2r\cos\beta_0,\\ y_{M1}\left(x,\tau_{M3}\left(x'\right)\right)-y_{M1}\left(x,\tau_{M2}\left(x'\right)\right) &\le s_3\left(r\right)\theta\cos\gamma_0.\end{aligned}$$

3. *The function*

$$y_R\left(x,t\right)=\left(y_{R1}\left(x,t\right),y_{R2}\left(x,t\right)\right)=\left(x\left(t\right),0\right)+z_R\left(x',t\right),\ t\ge 0,$$

is a unique solution to the Cauchy problem

$$y'=g_{D_R(r)}\left(y,x,x'\right),\ y\left(0\right)=\left(a,0\right)+\delta_r e\left(0\right),$$

in the whole interval $[0,\infty)$*, and satisfies the following conditions:*

$$y\left(x,t\right)-\left(x\left(t\right),0\right)\in \operatorname{bd}\, D_R\left(r\right),\ t\ge 0,$$

$$\begin{aligned} y_R\left(x,\tau_{R1}\left(x'\right)\right) &= \left(x\left(\tau_{R1}\left(x'\right)\right),0\right)+re\left(\beta_0\right),\ y_R\left(x,\tau_{R2}\left(x'\right)\right)\\ &= \left(x\left(\tau_{R2}\left(x'\right)\right),0\right)+re\left(\pi\right),\end{aligned}$$

$$\begin{aligned} y_{R1}\left(x,\tau_{M1}\left(x'\right)\right)-y_{R1}\left(x,0\right) &\le s_3\left(r\right)\theta\cos\gamma_0,\\ y_{M1}\left(x,\tau_{M2}\left(x'\right)\right)-y_{M1}\left(x,\tau_{M1}\left(x'\right)\right) &\le s_1\left(r\right)+r\left(1+\cos\beta_0\right).\end{aligned}$$

In the next two lemmas we analyze a situation when the invader passes, or attempts to pass, through the interval having two defenders on his left and right sides; see Figure 3.

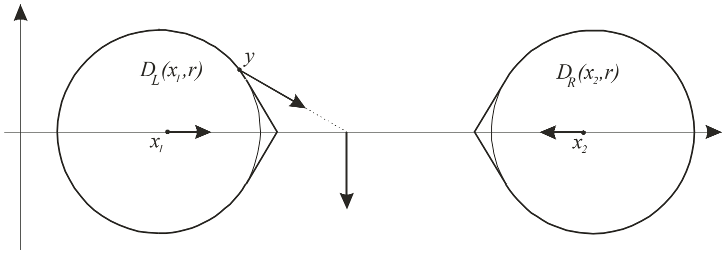


FIG. 3. *Invader attempts to pass between two defenders.*

LEMMA 4. *Given* $a,b\in\mathbb{R}$*, with* $|a-b|\ge\delta_r$*, and* $x\in X\left(a\right)$*, define*

$$y\left(t\right)=\left(b,-\theta t\right),\ t\ge 0.$$

Then we have

$$\left\|y\left(t\right)-\left(x\left(t\right),0\right)\right\|\ge r$$

for all $t\ge 0$.

Proof. The function x satisfies the Lipschitz condition with constant 1 so that

$$\begin{aligned}\|y(t)-(x(t),0)\|^2 &= (x(t)-b)^2+\theta^2t^2 \geq (|a-b|-t)^2+\theta^2t^2\\ &= \left(\theta^2+1\right)t^2-2|a-b|t+|a-b|^2.\end{aligned}$$

Since

$$\min_{t\geq 0}\left\{\left(\theta^2+1\right)t^2-2|a-b|t+|a-b|^2\right\}=|a-b|^2\frac{\theta^2}{\theta^2+1}$$

we obtain

$$\|y(t)-(x(t),0)\|\geq \delta_r\frac{\theta}{\sqrt{\theta^2+1}}=r$$

for every $t\geq 0$. □

LEMMA 5. *For all $x^-,\ x^+\in\mathbb{R}$ and $y=(y_1,y_2)\in\mathbb{R}^2$, with*

$$x^-\leq y_1\leq x^+,\ y_2\geq 0,$$

and for every $u\in[-1,1]$, we define

$$\xi\left(y,x^-,x^+\right)=\begin{cases} x^- & \textit{if} \quad y_1<\frac{1}{2}(x^-+x^+),\\ x^+ & \textit{if} \quad y_1\geq\frac{1}{2}(x^-+x^+)\end{cases}$$

and

$$g\left(y,x^-,x^+,u\right)=\begin{cases} g_{D_L(r)}\left(y,x^-,u\right) & \textit{if} \quad y_1<\frac{1}{2}\left(x^-+x^+\right),\\ g_{D_R(r)}\left(y,x^+,u\right) & \textit{if} \quad \frac{1}{2}\left(x^-+x^+\right)\leq y_1.\end{cases}$$

Suppose that

$$a^-,\ a^+\in\mathbb{R},\ a^-<a^+,\ b=\left(a^-,r\right)$$

and $x^-\in X\left(a^-\right)$, $x^+\in X\left(a^+\right)$ are chosen arbitrarily and satisfy the condition

$$x^-(t)\leq x^+(t),\ t\geq 0.$$

Then the Cauchy problem

$$y'=f\left(y,x^-(t),x^+(t),\xi'\left(y,x^-,x^+\right)\right),\ y(0)=b,$$

has a unique solution y defined in a maximal domain $[0,\tau(x^-,x^+)]$ satisfying the following conditions:

(a) $y_2(\tau(x^-,x^+))=0$ *or* $y_2(\tau(x^-,x^+))=r$.

(b)

$$x^-\left(\tau\left(x^-,x^+\right)\right)+\delta_r=y_1\left(\tau\left(x^-,x^+\right)\right)\leq x^+\left(\tau\left(x^-,x^+\right)\right)-\delta_r$$

and

$$y_1\left(\tau\left(x^-,x^+\right)\right)-y_1(0)\leq S_r(\pi-\beta_0)-S_r\left(\frac{\pi}{2}\right)+r\cos\beta_0+s_3(r)$$

whenever $y_2(\tau(x^-,x^+))=0$.

(c)

$$y_1\left(\tau\left(x^-,x^+\right)\right)=x^+\left(\tau\left(x^-,x^+\right)\right)$$

and

$$y_1\left(\tau\left(x^-,x^+\right)\right)-y_1(0)<s_2(r)+2r\cos\beta_0+2s_3(r)$$

when $y_2\left(\tau\left(x^-,x^+\right)\right)=r$.

(d)

$$\min\left\{\left\|y(t)-\left(x^-(t),0\right)\right\|,\left\|y(t)-\left(x^+(t),0\right)\right\|\right\}\geq r$$

for all $t\in\left[0,\tau\left(x^-,x^+\right)\right]$.

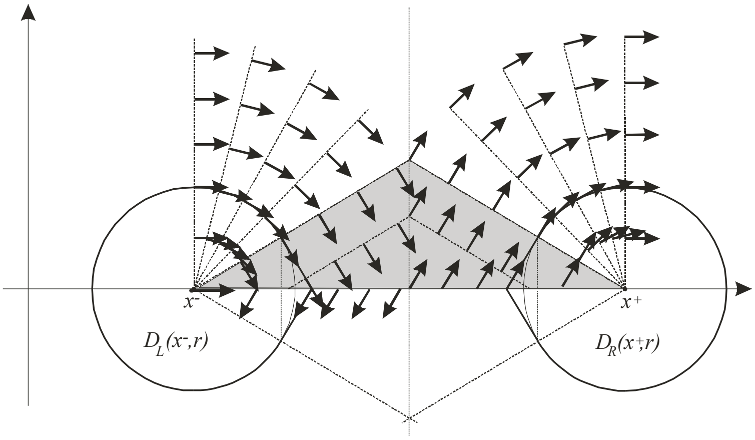


FIG. 4. *Vector field* g *with* $u=0$.

Proof. Note first (see Figure 4) that the domain of g is contained in $\mathbb{R}\times[0,\infty)$. Let us fix arbitrary $x^-\in X\left(a^-\right)$ and $x^+\in X\left(a^+\right)$. It follows from parts 1 and 3 of Corollary 1 that the Cauchy problem has a solution defined in a maximal domain $[0,\tau]$ satisfying conditions (a), (b), and (c). In view of part 2 or 3 of Corollary 1 the uniqueness follows from condition (d). Hence, it remains to prove condition (d). By part 1 of Corollary 1 we can define

$$t_1=\max\left\{t>0:\ y(t)\in\mathrm{bd}\ D_L\left(r,x^-(t)\right)\right\}.$$

Suppose $y_2(t_1)=0$ and $y_1(t_1)<\frac{1}{2}\left(x^-(t_1)+x^+(t_1)\right)$. Then $\tau\left(x^-,x^+\right)=t_1$, which completes the proof. In the opposite case we have

$$y_2(t_1)\geq 0,\ y_1(t_1)=\frac{1}{2}\left(x^-(t_1)+x^+(t_1)\right)$$

and

$$y(t_1)-\left(x^-(t_1),0\right)\in\mathrm{bd}\ D_L(r),\ y(t_1)-\left(x^+(t_1),0\right)\in\mathrm{bd}\ D_R(r).$$

Hence the two cases

$$\text{(d1)}\ \ h_r\leq y_2(t_1)\leq r\ \ \text{and}\ \ \text{(d2)}\ \ 0\leq y_2(t_1)<r$$

are possible; see Figure 5.

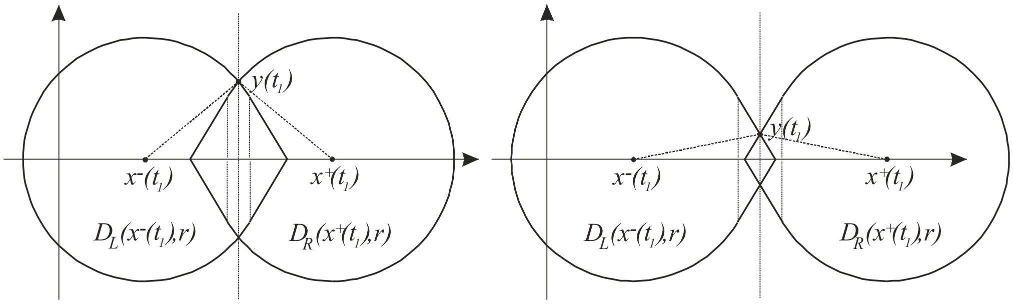


FIG. 5. *Cases* (d1) *and* (d2).

Case (d1). By part 3 of Corollary 1, there exists a $t_2 > t_1$ such that

$$y_2(t_2) = r \quad \text{and} \quad y_1(t_2) = x^+(t_2).$$

Clearly, we have $\tau(x^-, x^+) = t_2$. If we show that

$$y_1'(t) \geq \frac{1}{2}\Big(\big(x^-\big)'(t) + \big(x^+\big)'(t)\Big) \quad \text{a.e. in } [t_1, t_2],$$

then we will obtain

$$y_1(t) \geq \frac{1}{2}\big(x^-(t) + x^+(t)\big) \quad \text{in } [t_1, t_2].$$

This will obviously imply the needed estimate (d) because of the relations

$$\big\| y(t) - \big(x^-(t), 0\big)\big\| \geq \big\| y(t) - \big(x^+(t), 0\big)\big\| = r, \ t \in [t_1, t_2].$$

In order to simplify the notation we put

$$u_+ = \big(x^+\big)' \quad \text{and} \quad u_- = \big(x^-\big)'.$$

With this notation we have

$$\begin{aligned}
&2y_1'(t) - u_+(t) - u_-(t) \\
&= \frac{2y_2(t)}{r}\sqrt{\theta^2 - u_+^2(t)\frac{\left(x^+(t) - y_1(t)\right)^2}{r^2}} \\
&\quad + 2u_+(t)\frac{\left(x^+(t) - y_1(t)\right)^2}{r^2} - u_+(t) - u_-(t) \\
&= \frac{2y_2(t)}{r}\sqrt{\theta^2 - u_+^2(t)\frac{r^2 - y_2^2(t)}{r^2}} \\
&\quad + 2u_+(t)\frac{r^2 - y_2^2(t)}{r^2} - u_+(t) - u_-(t) \\
&= \frac{2y_2(t)}{r}\sqrt{\theta^2 - u_+^2(t)\left(1 - \frac{y_2^2(t)}{r^2}\right)} + u_+(t)\left(1 - \frac{2y_2^2(t)}{r^2}\right) - u_-(t) \\
&\geq \frac{2y_2(t)}{r}\sqrt{\theta^2 - u_+^2(t)\left(1 - \frac{y_2^2(t)}{r^2}\right)} + u_+(t)\left(1 - \frac{2y_2^2(t)}{r^2}\right) - 1
\end{aligned}$$

for almost all $t \in [t_1, t_2]$, since

$$|u_+(t)| \leq 1 \quad \text{and} \quad |u_-(t)| \leq 1$$

almost everywhere in $[0, \infty)$. Let us observe now that the function y_2 is increasing in the interval $[t_1, t_2]$, so that

$$\sin \beta_0 = \frac{h_r}{r} \leq \frac{y_2(t)}{r} \leq 1, \ t \in [t_1, t_2].$$

It is not very hard to verify that

$$\min_{(u,\zeta) \in [-1,1] \times [\sin \beta_0, 1]} \left\{ 2\zeta \sqrt{\theta^2 - u^2 (1 - \zeta^2)} + u \left(1 - 2\zeta^2\right) \right\} = 1.$$

Thus

$$2y_1'(t) - u_+(t) - u_-(t) \geq 0$$

almost everywhere in $[t_1, t_2]$, as claimed.

Case (d2). By the definition of g and by part 3 of Corollary 1, there exists a $t_2 > t_1$ such that $y_2(t_2) = h_r$. Moreover,

$$y'(t) = \left\langle \left(\left(x^+\right)'(t), 0 \right), Rw^+ \right\rangle Rw^+ + \sqrt{\theta^2 - \left\langle \left((x^+)'(t), 0 \right), Rw^+ \right\rangle^2} w^+$$

for almost all $t \in [t_1, t_2]$. Similarly to case (d1) it is enough to show that

$$2y_1'(t) - \left(x^+\right)'(t) - \left(x^-\right)'(t) \geq 0$$

for almost all $t \in [t_1, t_2]$, since at the end of this interval we arrive at the situation

$$y_1(t_2) \geq \frac{1}{2} \left(x^-(t_2) + x^+(t_2) \right), \ y(t_2) - \left(x^+(t_2), 0 \right) \in \operatorname{bd} D_R(r),$$

which can be treated in the same way as in case (d1).

With the notation used in case (d1), for almost all $t \in [t_1, t_2]$, we have

$$\begin{aligned}
&2y_1'(t) - \left(x^+\right)'(t) - \left(x^-\right)'(t) \\
&= 2\sqrt{\theta^2 - u_+^2(t) \cos^2 \beta_0} \sin \beta_0 + 2u_+(t) \cos^2 \beta_0 - u_+(t) - u_-(t) \\
&= 2\sqrt{\theta^2 - u_+^2(t) \cos^2 \beta_0} \sin \beta_0 + u_+(t) \cos 2\beta_0 - u_-(t) \\
&\geq 2\sqrt{\theta^2 - u_+^2(t) \cos^2 \beta_0} \sin \beta_0 + u_+(t) \cos 2\beta_0 - 1 \\
&\geq 2\sqrt{\theta^2 - \cos^2 \beta_0} \sin \beta_0 - \cos 2\beta_0 - 1 \\
&= 0,
\end{aligned}$$

because of the relation $\beta_0 \in \left(0, \frac{\pi}{4}\right)$. The proof of Lemma 5 is complete. □

We are now ready to define an invasion strategy. The strategy will be constructed with the aid of a vector field

$$g : \mathbb{R} \times [0, \infty) \times \mathbb{R}^n \times [-1, 1] \to \mathbb{R}^2.$$

In the half-plane $\mathbb{R} \times (0, \infty)$, the invader's trajectory will be determined by g in such a way that he will move right with maximal speed along the boundary of the union

$$D_L(x_1(t), \varrho) \cup \bigcup_{k=2}^{n-1} D_M(x_k(t), \varrho) \cup D_R(x_n(t), \varrho).$$

Since defender trajectories can intersect, we have to introduce a few auxiliary notions.

For

$$x = (x_1, \ldots, x_n) \in \mathbb{R}^n \text{ and } y = (y_1, y_2) \in \mathbb{R}^2,$$

we put

$$J^-(x, y) = \{j \in \{1, 2, \ldots, n\} : x_j \le y_1\}, \ J^+(x, y) = \{j \in \{1, 2, \ldots, n\} : x_j > y_1\},$$

$$\xi^-(x, y) = \begin{cases} \max_{j \in J^-(x,y)} x_j, & \text{when } J^-(x, y) \ne \varnothing, \\ \min\{x_1, \ldots, x_n\}, & \text{when } J^-(x, y) = \varnothing, \end{cases}$$

$$\xi^+(x, y) = \begin{cases} \min_{j \in J^-(x,y)} x_j, & \text{when } J^+(x, y) \ne \varnothing, \\ \max\{x_1, \ldots, x_n\}, & \text{when } J^+(x, y) = \varnothing. \end{cases}$$

Let us define, for arbitrary

$$y = (y_1, y_2) \in \mathbb{R} \times [0, \infty), \ x = (x_1, \ldots, x_n) \in \mathbb{R}^n, \ u \in [-1, 1],$$

$$\xi(x, y) = \begin{cases} \xi^-(x, y), & \text{when } y_1 < \frac{1}{2}(\xi^-(x, y) + \xi^+(x, y)), \\ \xi^+(x, y), & \text{when } y_1 \ge \frac{1}{2}(\xi^-(x, y) + \xi^+(x, y)) \end{cases}$$

and

$$g(y, x, u) = \begin{cases} g_{D_L(\varrho)}(y, \xi(x, y), u), & \text{when } y_1 < \frac{1}{2}(\xi^-(x, y) + \xi^+(x, y)), \\ g_{D_R(\varrho)}(y, \xi(x, y), u), & \text{when } y_1 \ge \frac{1}{2}(\xi^-(x, y) + \xi^+(x, y)). \end{cases}$$

Let us choose arbitrary $a_1, \ldots, a_n \in \mathbb{R}$, with

$$a_1 \le a_2 \le \cdots \le a_n, \ a_1 \le r,$$

and take

$$b = (a_1 - \varrho, 0).$$

We define now an invasion strategy

$$\sigma_{\mathbf{I}} : X(a_1, \ldots, a_n) \to Y(b).$$

Let us fix arbitrary $x = (x_1, \ldots, x_n) \in X(a_1, \ldots, a_n)$. By Corollary 1, the Cauchy problem

$$y' = g(y, x, \xi'(x, y)), \ y(0) = b,$$

has a unique solution $\eta(x, \cdot) = (\eta_1(x, \cdot), \eta_2(x, \cdot))$ defined in a maximal interval $[0, \tau(x)]$ and such that

$$\eta_2(x, \tau(x)) = 0.$$

For all $t \geq 0$ we put

$$\sigma_{\mathbf{I}}(y)(t) = \begin{cases} \eta(x,t) & \text{if} \quad 0 \leq t \leq \tau(x), \\ \eta(x,\tau(x)) + (t-\tau(x))(0,-\theta) & \text{if} \quad t > \tau(x),\ J^{+}(x(\tau(x)),\eta(x,\tau(x))) \neq \varnothing, \\ \eta(x,\tau(x)) + (t-\tau(x))(\theta,0) & \text{if} \quad t > \tau(x),\ J^{+}(x(\tau(x)),\eta(x,\tau(x))) = \varnothing. \end{cases}$$

It is easy to see that $\sigma_{\mathbf{I}} : X(a_1,\ldots,a_n) \to Y(b)$ is a nonanticipating function.

PROPOSITION 1. *Strategy $\sigma_{\mathbf{I}}$ rushes for the line segment $[0,\Delta(\varrho)] \times \{0\}$ successfully.*

Proof. Let $x \in X(a_1,\ldots,a_n)$ be chosen arbitrarily. Take $y = \sigma_{\mathbf{I}}(x)$ and $\tau(x) = t^*$. With the aid of Corollary 1 and Lemma 5 we get the estimate

$$\min_{j=1,2,\ldots,n} \|y(t) - x_j(t)\| \geq \varrho,\ t \in [0,t^*],$$

and the relation

$$J^{-}(x(t^*),y(t^*)) \neq \varnothing.$$

By the definition of $\sigma_{\mathbf{I}}$, if $J^{+}(x(t^*),y(t^*)) = \varnothing$, then

$$\min_{j=1,2,\ldots,n} \|y(t) - x_j(t)\| \geq \varrho,\ t \geq t^*.$$

In the case of $J^{+}(x(t^*),y(t^*)) \neq \varnothing$ we have

$$\xi^{-}(x(t^*),y(t^*)) + \delta_r = y_1(t^*) \leq \xi^{+}(x(t^*),y(t^*)) + \delta_r,$$

which in view of Lemma 4 gives

$$\min_{j=1,2,\ldots,n} \|y(t) - x_j(t)\| \geq \varrho,\ t \geq t^*.$$

Applying Corollary 1 and Lemma 5 once more we obtain

$$y_1(t^*) - y_1(0) < 2\Lambda_1(r) + (n-2)\Lambda_1(r) = \Delta(\varrho),$$

which finishes the proof. □

Conclusion. We have just considered a special initial situation (a,b). Since the invader has an advantage in velocity he can always reach such a situation. Therefore defenders are not able to defend the interval $[0,\Delta(\varrho)] \times \{0\}$.

4. Defense. We will consider here the case of $n \geq 3$. The case $n = 1$ is easier, and the case $n = 2$ can be solved similarly. Let us fix an $r \in (0,\varrho)$. A guarded interval of the form $[0,\Delta(r)] \times \{0\}$ will be divided into n subintervals. Let y be a trajectory of the invader. Each defender will guard his own subinterval until, for a $t^* > 0$ and a $k > 1$,

$$x_{k-1}(t^*) \leq y_1(t^*) \leq x_k(t^*) \quad \text{and} \quad y_2(t^*) = 0.$$

If such a situation happens, then defenders D_{k-1} and D_k will apply trajectories

$$x_{k-1}(t) = x_{k-1}(t^*) + t - t^* \text{ and } x_k(t) = x_k(t^*) - t + t^*,$$

respectively. In other words, defenders do not cooperate except when the invader enters the interval at a point placed between two of them. The situation described above will be analyzed from the point of view of the defenders in the following lemma.

LEMMA 6. *Given a^-, $a^+ \in \mathbb{R}$, $b = (b_1, 0) \in \mathbb{R}^2$ satisfying the conditions*

$$a_1^- < b_1 < a_1^+ \quad \textit{and} \quad a_1^+ - a_1^- \leq 2\delta_r,$$

define

$$x^-(t) = \left(a^- + t, 0\right), \ x^+(t) = \left(a^+ - t, 0\right), \ t \geq 0.$$

Then, for each $y \in Y(b)$, there exists a $t_0 \geq 0$ such that

$$\min\left\{\left\|y(t_0) - \left(x^-(t_0), 0\right)\right\|, \left\|y(t_0) - \left(x^+(t_0), 0\right)\right\|\right\} \leq r.$$

Proof. Without loss of the generality we may assume that

$$a^- + r < b_1 < a^+ - r.$$

Let us take

$$t_1 = \frac{h_r}{\theta}.$$

We have

$$x^+(t_1) = a^+ - t_1 - r\cos\beta_0 \leq a^- + t_1 + r\cos\beta_0 = x^-(t_1),$$

because of the equalities

$$2t_1 + 2r\cos\beta_0 = 2r\left(\frac{\sin\beta_0}{\theta} + \cos\beta_0\right) = \frac{2r}{\cos\beta_0} = 2\delta_r.$$

Thus, there exists a $t_0 \leq t_1$ such that

$$x^-(t_0) + r\cos\beta_0 = y_1(t_0) \quad \text{or} \quad x^+(t_0) - r\cos\beta_0 = y_1(t_0).$$

Since

$$|y_2(t_0)| \leq \theta t_0 \leq h_r,$$

we obtain

$$\left\|y(t_0) - \left(x^-(t_0), 0\right)\right\|^2 \leq r^2 \quad \text{or} \quad \left\|y(t_0) - \left(x^+(t_0), 0\right)\right\|^2 \leq r^2,$$

which completes the proof. □

Note that in the situation analyzed in Lemma 6, the invader loses the game because of the inequality $r < \varrho$. We are now going to describe the subintervals mentioned above and suitable guarding strategies.

For each $r \in (0, \varrho)$ and every $k = 1, 2, \ldots, n$, we put

$$\begin{aligned}
&\Delta_1(r) = \Lambda_1(r), \\
&\Delta_k(r) = \Delta_1(r) + (k-1)\Lambda_2(r), \ k = 2, 3, \ldots, n-1, \\
&\Delta_n(r) = 2\Lambda_1(r) + (n-2)\Lambda_2(r).
\end{aligned}$$

Defender $\mathbf{D}_k$ will guard the interval

$$[\Delta_{k-1}(r), \Delta_k(r)] \times \{0\},$$

where

$$\Delta_0(r) \overset{\text{def}}{=} 0.$$

We are now going to describe guarding strategies. In order to do this we will divide the whole plane $\mathbb{R}^2$ into suitable regions. First of all we put (see Corollary 1)

$$\begin{aligned} a_1^+(r) &= \Delta_1(r) - \varkappa_r,\ c_1^+(r) = \Delta_1(r) - \varkappa_r - \frac{1}{2}\delta_r, \\ x_1(t) &= \begin{cases} r+t & \text{if} \quad 0 \le t < \tau_{L2}(x_1'), \\ x_1(\tau_{L2}(x_1')) - t + \tau_{L2}(x_1') & \text{if} \quad \tau_{L2}(x_1') \le t \le 2\tau_{L2}(x_1'), \end{cases} \end{aligned}$$

$$\begin{aligned} a_k^-(r) &= \Delta_{k-1}(r) + \varkappa_r,\ c_k^-(r) = \Delta_{k-1}(r) + \varkappa_r + \frac{1}{2}\delta_r, \\ a_1^+(r) &= \Delta_k(r) - \varkappa_r,\ c_k^+(r) = \Delta_k(r) - \varkappa_r - \frac{1}{2}\delta_r, \\ x_k(t) &= \begin{cases} \Delta_{k-1} + \delta_r + t & \text{if} \quad 0 \le t < \tau_{M3}(x_k'), \\ x_k(\tau_{M3}(x_k')) - t + \tau_{M3}(x_k') & \text{if} \quad \tau_{M3}(x_k') \le t \le 2\tau_{M3}(x_k') \end{cases} \end{aligned}$$

for $k = 2, 3, \ldots, n-1$, and

$$\begin{aligned} a_n^-(r) &= \Delta_{n-1}(r) + \varkappa_r,\ c_n^-(r) = \Delta_{n-1}(r) + \varkappa_r + \frac{1}{2}\delta_r, \\ x_n(t) &= \begin{cases} \Delta_{n-1} + \delta_r + t & \text{if} \quad 0 \le t < \tau_{R2}(x_n'), \\ x_n(\tau_{R2}(x_n')) - t + \tau_{R2}(x_n') & \text{if} \quad \tau_{R2}(x_n') \le t \le 2\tau_{R2}(x_n'). \end{cases} \end{aligned}$$

Partition 1. Set (see Corollary 1)

$$\begin{aligned} \Omega_1(r) &= \operatorname{conv} y_L(x_1, [0, 2\tau_{L,2}(r)]), \\ C_{R1}(r) &= \left\{(y_1, y_2) \in \mathbb{R}^2 : \ |y_2| < \left(y_1 - a_1^+(r)\right)\cot\gamma_0\right\}, \\ C_1(r) &= \mathbb{R}^2 \diagdown C_{R1},\ \Omega_1^0(r) = \operatorname{int}(\Omega_1(r) \diagdown C_{R1}). \end{aligned}$$

Figure 6 illustrates the partition 1, with $r = 1$ and $\theta = \sqrt{2}$.

Partition k. Set (see Corollary 1)

$$\begin{aligned} \Omega_k(r) &= \operatorname{conv} y_M(x_k, [0, 2\tau_{M,3}(r)]), \\ C_{Lk}(r) &= \left\{(y_1, y_2) \in \mathbb{R}^2 : \ |y_2| < \left(a_{k-1}^+(r) - y_1\right)\cot\gamma_0\right\}, \\ C_{Rk}(r) &= \left\{(y_1, y_2) \in \mathbb{R}^2 : \ |y_2| < \left(y_1 - a_k^+(r)\right)\cot\gamma_0\right\}, \\ C_k(r) &= \mathbb{R}^2 \diagdown (C_{Lk} \cup C_{Rk}),\ \Omega_k^0(r) = \operatorname{int}(\Omega_1(r) \diagdown (C_{Lk} \cup C_{Rk})) \end{aligned}$$

for $k = 2, 3, \ldots, n-1$. Figure 7 illustrates the partition k, with $r = 1$ and $\theta = \sqrt{2}$.

Partition n. Set (see Corollary 1)

$$\begin{aligned} \Omega_n(r) &= \operatorname{conv} y_R(x_n, [0, 2\tau_{R,2}(r)]), \\ C_{Ln}(r) &= \left\{(y_1, y_2) \in \mathbb{R}^2 : \ |y_2| < \left(a_{n-1}^+(r) - y_1\right)\cot\gamma_0\right\}, \\ C_n(r) &= \mathbb{R}^2 \diagdown C_{Ln},\ \Omega_n^0(r) = \operatorname{int}(\Omega_n(r) \diagdown C_{Rk}). \end{aligned}$$

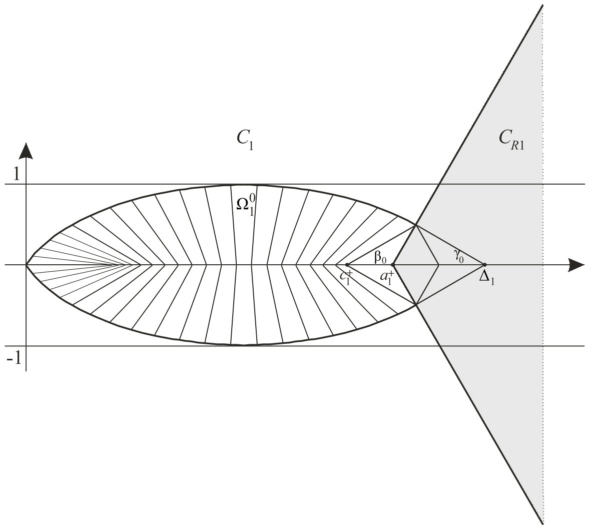


FIG. 6. *Partition* Π_1.

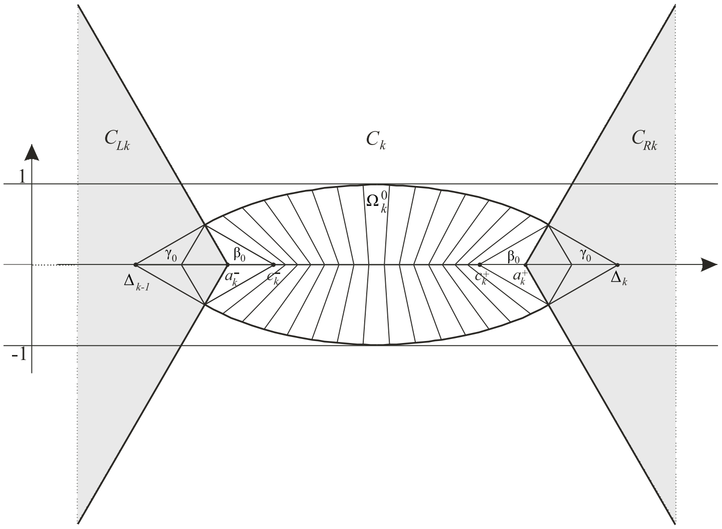


FIG. 7. *Partition* Π_k*, for* $1 < k < n - 1$.

Figure 8 illustrates the partition n, with $r = 1$ and $\theta = \sqrt{2}$.

Before we give the definition of guarding strategy, we have to answer some questions. For example, consider Partition k, with $k \in \{2, 3, \ldots, n-1\}$. Take arbitrary

$$b \in \mathbb{R}^2 \diagdown \text{int } \left(C_k\left(r\right) \cap \Omega_k\left(r\right)\right) \quad \text{and } y \in Y\left(b\right).$$

If $z \in C_k\left(r\right) \diagdown$ int $\Omega_k\left(r\right)$, then there exist exactly one $\Pi_{\Omega_k(r)}\left(z\right) \in$ bd $\Omega_k\left(r\right)$ such that

$$\left\| z - \Pi_{\Omega_k(r)}\left(z\right) \right\| = \text{dist}\left(z, \Omega_k\left(r\right)\right)$$

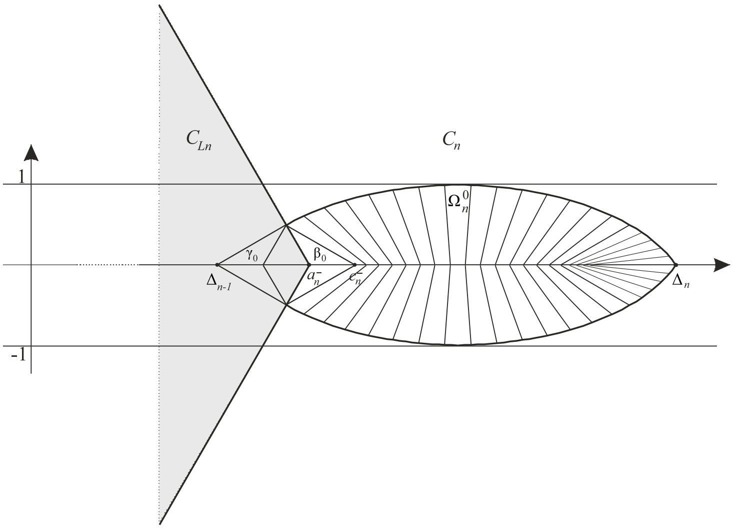


FIG. 8. *Partition* Π_π.

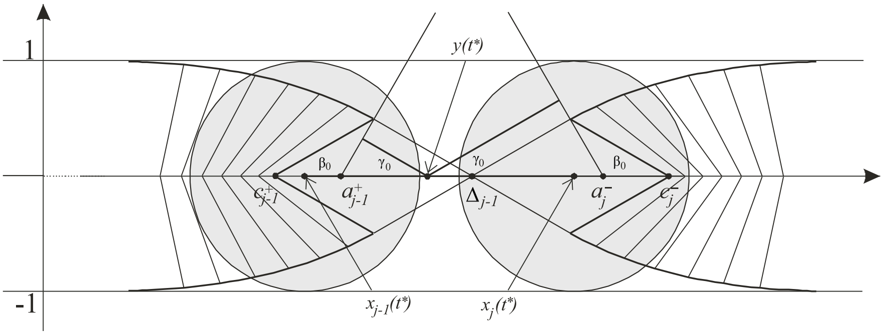


FIG. 9. *Invader will be captured.*

and exactly one $\varphi_k(z) \in [\beta_0, \pi - \beta_0] \cup [\pi - \beta_0, 2\pi)$ such that

$$\Pi_{\Omega_k(r)}(z) = y_M(x_k, \varphi_k(z)).$$

If $z \in C_{Lk}(r)$, then, for

$$f_{Lk}(z) \overset{\text{def}}{=} c_k^-(r) + \frac{1}{\theta}\text{dist}(z, \text{bd } C_{Lk}(r)),$$

we have

$$\text{dist}((f_{Lk}(z), 0), \text{bd } C_{Lk}(r)) = \frac{1}{\theta}\text{dist}(z, \text{bd } C_{Lk}(r)).$$

See Figure 9.

Similarly, if $z \in C_{Rk}(r)$, then, for

$$f_{Rk}(z) \overset{\text{def}}{=} c_k^+(r) - \frac{1}{\theta}\text{dist}(z, \text{bd } C_{Rk}(r)),$$

we have

$$\operatorname{dist}\left(\left(f_{Rk}\left(z\right),0\right),\operatorname{bd}\,C_{Rk}\left(r\right)\right)=\frac{1}{\theta}\operatorname{dist}\left(z,\operatorname{bd}\,C_{Rk}\left(r\right)\right).$$

Finally, note that

$$\begin{aligned}f_{Lk}\left(z\right)&=y_M\left(x_k,\varphi_k\left(z\right)\right)=\left(x_k\left(\varphi_k\left(z\right)\right),0\right)+re\left(\varphi_k\left(z\right)\right)\text{ if }z\in\operatorname{bd}\,C_{Lk}\left(r\right),\\ f_{Rk}\left(z\right)&=y_M\left(x_k,\varphi_k\left(z\right)\right)=\left(x_k\left(\varphi_k\left(z\right)\right),0\right)+re\left(\varphi_k\left(z\right)\right)\text{ if }z\in\operatorname{bd}\,C_{Rk}\left(r\right).\end{aligned}$$

These observations allow us to define a Lipschitzian function $f_k:\mathbb{R}^2\diagdown\Omega_k^0\left(r\right)\to\mathbb{R}$ by the formula

$$f_k\left(z\right)=\begin{cases}f_{Lk}\left(z\right), & \text{when}\quad z\in C_{Lk}\left(r\right),\\ x_k\left(\varphi_k\left(z\right)\right), & \text{when}\quad z\in C_k\left(r\right)\diagdown\Omega_k^0\left(r\right),\\ f_{Rk}\left(z\right), & \text{when}\quad z\in C_{Rk}\left(r\right).\end{cases}$$

Suppose that, for an interval $[t_1,t_2]$, we have

$$y\left(t\right)\in C_k\left(r\right)\diagdown\operatorname{int}\,\Omega_k\left(r\right),\ t\in\left[t_1,t_2\right].$$

Taking into account the fact that $\Pi_{\Omega_k(r)}$ is a nonexpansive mapping, one can show that

$$\left|x_k'\left(\varphi_k\left(y\left(t\right)\right)\right)\right|\leq\frac{1}{\theta}\left\|y_M'\left(x_k,\varphi_k\left(y\left(t\right)\right)\right)\right\|\leq1$$

almost everywhere in $\left[t_1,t_2\right]$. Suppose now that, for an interval (t_1,t_2), we have

$$y\left(t\right)\in C_{Lk}\left(r\right),\ t\in\left(t_1,t_2\right).$$

Since the function $\operatorname{dist}(\cdot,\operatorname{bd}\,C_{Lk}\left(r\right))$ satisfies the Lipschitz condition with constant 1, we have

$$\left|\frac{d}{dt}f_{Lk}\left(y\left(t\right)\right)\right|\leq\frac{1}{\theta}\left\|y'\left(t\right)\right\|\leq1$$

almost everywhere in $\left(t_1,t_2\right)$. Clearly, the analogous statement is true for the function f_{Rk}. It follows from the considerations above that

$$\left|\frac{d}{dt}f_k\left(y\left(t\right)\right)\right|\leq1$$

for almost all $t\in E_k\left(y\right)$, where

$$E_k\left(y\right)=\left\{t\geq0:\ y\left(t\right)\notin\Omega_k^0\left(r\right)\right\}.$$

Repeating the construction described for Partitions 1 and n, we obtain functions

$$f_1:\mathbb{R}^2\diagdown\Omega_1^0\left(r\right)\to\mathbb{R},\ f_n:\mathbb{R}^2\diagdown\Omega_n^0\left(r\right)\to\mathbb{R}$$

and multifunctions

$$E_1:Y\left(b\right)\to\left[0,\infty\right),\quad E_n:Y\left(b\right)\to\left[0,\infty\right)$$

with analogous properties.

Now we are ready to define a guarding strategy. Let

$$b = (b_1, b_2) \in \mathbb{R}^2 \diagdown \bigcup_{k=1}^{n} \text{int } \Omega_k (r),$$

and let $y \in Y(b)$ be chosen arbitrarily. We define

$$a_k = f_k (b), \ k = 1, 2, \ldots, n.$$

Clearly, we may assume that

$$\| b - (a_k, 0) \| > r, \ k = 1, 2, \ldots, n.$$

The following two cases are possible:

(a) There exist a $t^* > 0$ and a $k \in \{1, 2, \ldots, n\}$ such that $\| y(t) - (f_k (y(t)), 0) \| \leq r$.

(b) $\| y(t) - (f_k (y(t)), 0) \| > r$ for all $t \geq 0$ and every $k = 1, 2, \ldots, n$.

In case (a) we take

$$\tau_1^* (y) = \inf \left\{ t > 0 : \min_{k=1,2,\ldots,n} \| y(t) - (f_k (y(t)), 0) \| > r \right\}$$

and define, for every $k = 1, 2, \ldots, n$,

$$\sigma_{\mathbf{D}_k} (y)(t) = \begin{cases} f_k (y(t)) & \text{for} \quad t \in [0, \tau_1^* (y)], \\ f_k (y(\tau_1^* (y))) & \text{for} \quad t \geq \tau_1^* (y), \end{cases}$$

$$\sigma_{\mathbf{D}} (y)(t) = (\sigma_{\mathbf{D}_1} (y)(t), \ldots, \sigma_{\mathbf{D}_n} (y)(t)), \ t \geq 0.$$

In case (b) it may be

(b1) $y(t) \notin [0, \Delta_n (r)] \times \{0\}$ for all $t \geq 0$, or

(b2) there exists a $t^* > 0$ such that $y(t^*) \in [0, \Delta_n (r)] \times \{0\}$.

If case (b1) takes place, then we define, for every $k = 1, 2, \ldots, n$,

$$\sigma_{\mathbf{D}_k} (y)(t) = f_k (y(t)), \ t \geq 0,$$

$$\sigma_{\mathbf{D}} (y)(t) = (\sigma_{\mathbf{D}_1} (y)(t), \ldots, \sigma_{\mathbf{D}_n} (y)(t)), \ t \geq 0.$$

In case (b2) we take

$$\tau_2^* (y) = \inf \{ t > 0 : \ y(t) \notin [0, \Delta_n (r)] \times \{0\} \}.$$

It is easy to see that it must be $y_1 (\tau_2^* (y)) = 0$ and

$$c_{j-1}^{+} < y_1 (\tau_2^* (y)) < c_j^{-}$$

for a $j \in \{2, 3, \ldots, n\}$. Taking into account Lemma 6 we define

$$\sigma_{\mathbf{D}_{j-1}} (y)(t) = \begin{cases} f_{j-1} (y(t)) & \text{for} \quad t \in [0, \tau_2^* (y)], \\ f_{j-1} (y(\tau_2^* (y))) + t - \tau_2^* (y) & \text{for} \quad t \geq \tau_2^* (y), \end{cases}$$

$$\sigma_{\mathbf{D}_j} (y)(t) = \begin{cases} f_j (y(t)) & \text{for} \quad t \in [0, \tau_2^* (y)], \\ f_j (y(\tau_2^* (y))) - t + \tau_2^* (y) & \text{for} \quad t \geq \tau_2^* (y), \end{cases}$$

$$\sigma_{\mathbf{D}_k} (y)(t) = \begin{cases} f_k (y(t)) & \text{for} \quad t \in [0, \tau_2^* (y)], \\ f_k (y(\tau_2^* (y))) & \text{for} \quad t \geq \tau_2^* (y) \end{cases}$$

for $k \in \{1,2,\dots,n\} \diagdown \{j-1,j\}$, and finally

$$\sigma_{\mathbf{D}}(y)(t) = (\sigma_{\mathbf{D}_1}(y)(t),\dots,\sigma_{\mathbf{D}_n}(y)(t)),\ t \geq 0.$$

In this way we have defined a nonanticipating mapping $\sigma_{\mathbf{D}}$ satisfying the inequalities

$$\left|\frac{d}{dt}\sigma_{\mathbf{D}_k}(y)(t)\right| \leq 1,\ k = 1,2,\dots,n,$$

almost everywhere in $[0,\infty)$. By virtue of what has been said,

$$\sigma_{\mathbf{D}} : Y(b) \to X(a_1,\dots,a_n)$$

is a guarding strategy.

Proposition 2. *The strategy $\sigma_{\mathbf{D}}$ guards the interval $[0,\Delta_n(r)] \times \{0\}$ successfully.*

Proof. Let us fix a $y \in Y(b)$ and put

$$x = (x_1,\dots,x_n) = \sigma_{\mathbf{D}}(y).$$

By the definition of $\sigma_{\mathbf{D}}$ it is enough to consider the case when, for a $t^* > 0$,

$$y(t^*) \in (0,\Delta_n(r)) \times \{0\}.$$

Without loss of generality we may assume that

$$y(t) \notin (0,\Delta_n(r)) \times \{0\},\ t \in [0,t^*),$$

and

$$\min_{k=1,2,\dots,n} \|y(t) - (x_k(t),0)\| > r,\ t \in [0,t^*].$$

Clearly, $y_2(t^*) = 0$ and there exists a $j \in \{2,3,\dots,n\}$ such that

$$c_{j-1}^{+} < y_1(t^*) < c_j^{-}.$$

We claim that

$$\min\{\|y(t) - (x_{j-1}(s^*),0)\|, \|y(t) - (x_j(s^*),0)\|\} \leq r$$

for an $s^* > t^*$. In view of Lemma 6, it suffices to check the inequality

$$x_j(t^*) - x_{j-1}(t^*) \leq 2\delta_r.$$

By the definition of the strategy $\sigma_{\mathbf{D}}$ we have (see Figure 9)

$$x_{j-1}(t^*) = c_{j-1}^{+} + d_1 < y_1(t^*) < c_j^{-} - d_2 = x_j(t^*),$$

where

$$d_1 = \frac{1}{v}\operatorname{dist}\left(y(t^*), B_{j-1}^{+}(r)\right) \quad \text{and} \quad d_1 = \frac{1}{v}\operatorname{dist}\left(y(t^*), B_j^{-}(r)\right).$$

Since

$$\operatorname{dist}\left(y(t^*), B_{j-1}^{+}(r)\right) = \left(y_1(t^*) - a_{j-1}^{+}\right)\cos\gamma_0,$$
$$\operatorname{dist}\left(y(t^*), B_j^{-}(r)\right) = \left(a_j^{-} - y_1(t^*)\right)\cos\gamma_0$$

we have

$$\begin{aligned}x_j(t^*) - x_{j-1}(t^*) &= c_j^- - c_{j-1}^+ - (d_1 + d_2)\\ &= \Delta_{j-1}(r) + \varkappa_r + \frac{1}{2}\delta_r - \left(\Delta_{j-1}(r) - \varkappa_r - \frac{1}{2}\delta_r\right)\\ &\quad -\frac{1}{v}\left(a_j^- - a_{j-1}^+\right)\cos\gamma_0\\ &= 2\varkappa_r + \delta_r - \frac{2\varkappa_r}{v}\cos\gamma_0\\ &= 2\delta_r.\end{aligned}$$

Therefore, if

$$\min_{k=1,2,\ldots,n} \|y(t) - (x_k(t), 0)\| > r,\ t \geq 0,$$

then

$$y(t^*) \notin [0, \Delta_n(r)] \times \{0\},\ t \geq 0,$$

which means that $\sigma_{\mathbf{D}}$ guards the interval $[0, \Delta_n(r)] \times \{0\}$ successfully and completes the proof. □

Conclusion. Theorem 1 follows now from Propositions 1 and 2 and the equality

$$\lim_{r \uparrow \varrho} \Delta_n(r) = \Delta(\varrho).$$

5. Conclusions. 1. In cases $n = 1$ and $n = 2$ we have

$$L_1 \stackrel{\text{def}}{=} \Delta(\varrho) = \frac{r}{\theta^2 - 1}\int_0^\pi \left(\sqrt{\theta^2 - \cos^2\alpha} + \sin\alpha\right) d\alpha$$

and

$$\begin{aligned}L_2 \stackrel{\text{def}}{=} \Delta(\varrho) &= \frac{2r}{\theta^2 - 1}\int_0^{\pi - \beta_0} \left(\sqrt{\theta^2 - \cos^2\alpha} + \sin\alpha\right) d\alpha\\ &\quad + \frac{2r\theta}{(\theta^2 - 1)\sqrt{\theta^2 + 1}} + \frac{2r\theta}{\sqrt{\theta^2 + 1}},\end{aligned}$$

respectively. It is not very hard to verify that $L_2 > 2L_1$, which is not a surprise. This means that the proper cooperation gives a benefit since $2L_1$ is the maximal length of an interval guarded by two defenders separately.

2. It was assumed in the paper that motion of the defenders is restricted to a straight line. One can show, without such a restriction, that the maximal length of a guarded interval is then greater than $\Delta(\varrho)$ (which is also not a surprise). Unfortunately we are not able to estimate the maximal length of the guarded interval.

3. It seems that guarding-line-segment games can be used as a tool in solving pursuit-evasion games in various environments. For example, it can be shown that optimal strategies of “wall pursuit” game (see Example 9.5.2 in [8]) coincide in an adequate region with strategies of guarding-line-segment games with one defender, when its motion is not restricted to a straight line. It gives an answer to the question raised in Problem 9.5.1 of [8].

Acknowledgment. I would like to thanks the anonymous reviewers for their helpful comments.

REFERENCES

[1] N. N. Krasovskii and A. I. Subbotin, *Positional Differential Games*, Nauka, Moscow, 1974 (in Russian).

[2] C. Ryll-Nardzewski, *A theory of pursuit and evasion*, Ann. Math. Stud., 3 (1957), pp. 393–405.

[3] J. O'Rourke, *Art Gallery Theorems and Algorithms*, Oxford University Press, Oxford, 1987.

[4] R. Murietta-Cid, A. Sarmiento, S. Battacharya, and S. Hutchinson, *Maintaining visibility of a moving target at a fixed distance*, in Proceedings of the IEEE International Conference on Robotics and Automation, 2004, pp. 479–484.

[5] V. Isler, S. Kannan, and S. Khanna, *Locating and Capturing an Evader in a Polygonal Environment*, in Proceedings of the 6th International Workshop on the Algorithmic Foundations of Robotics (WAFR'04), Springer Tracts Adv. Robotics 17, Springer-Verlag, Berlin, 2005, pp. 251–266.

[6] E. Klavins and M. Murray, *Distributed algorithms for cooperative control*, IEEE Pervasive Comput. 3 (2004), pp. 56–65.

[7] S. D. Bopardikar, F. Bullo, and J. Hespanha, *Sensing limitations in the Lion and Man problem*, in American Control Conference, 2007 (ACC '07), pp. 5958–5963.

[8] R. Isaacs, *Differential Games*, John Wiley and Sons, New York, London, Sydney, 1965.

[9] A. Chikrii, *Conflict-Controlled Processes*, Kluwer Academic, Dordrecht, Boston, London, 1977.

[10] W. Rzymowski and A. Stachura, *Solution to a problem of guarding territory*, Systems Control Lett., 7 (1986), pp. 71–72.

SIAM J. Control Optim.
Vol. 48, No. 2, pp. 839–858

© 2009 Society for Industrial and Applied Mathematics

FOCUSING WAVES IN UNKNOWN MEDIA BY MODIFIED TIME REVERSAL ITERATION[*]

MATIAS F. DAHL[†], ANNA KIRPICHNIKOVA[‡], and MATTI LASSAS[§]

Abstract. We study the wave equation on a bounded domain M in $\mathbb{R}^m$ or on a compact Riemannian manifold with boundary. Assume that we do not know the coefficients of the wave equation but are given only the hyperbolic Robin-to-Dirichlet map that corresponds to physical measurements on a part of the boundary. In this paper we show that at a fixed time t_0 a wave can be cut off outside a suitable set. That is, if $N \subset M$ is a union of balls in the travel time metric having centers at the boundary, then we can modify a given Robin boundary value of a wave such that at time t_0 the modified wave is arbitrarily close to the original wave inside N and arbitrarily small outside N. Also, at time t_0 the time derivative of the modified wave is arbitrarily small in all of M. We apply this result to construct a sequence of Robin boundary values so that at a time t_0 the corresponding waves converge to a delta distribution $\delta_{\widehat{x}}$ while the time derivative of the waves converge to zero. Such boundary values are generated by an iterative sequence of measurements. In each iteration step we apply time reversal and other simple operators to measured data and compute boundary values for the next iteration step. A key feature of this result is that it does not require knowledge of the coefficients in the wave equation, that is, of the material parameters inside the media. However, we assume that the point $\widehat{x}$ where the wave focuses is known in travel time coordinates, and $\widehat{x}$ satisfies a certain geometrical condition.

Key words. focusing of waves, wave equation, time reversal

AMS subject classifications. 35R30, 93B05

DOI. 10.1137/070705192

1. Introduction.

Let us consider the wave equation in M that is a bounded domain of $\mathbb{R}^m$ or a compact manifold:

$$\begin{cases} u_{tt}(x,t) + \mathcal{A}u(x,t) = 0 \quad \text{in } M \times \mathbb{R}_+, \\ u|_{t=0} = 0, \quad u_t|_{t=0} = 0, \\ (\partial_\nu - \sigma)u|_{\partial M \times \mathbb{R}_+} = f, \end{cases} \tag{1.1}$$

where A is a second order elliptic partial differential operator, σ is a smooth function $\sigma \in C^\infty(\partial M)$, and $f \in L^2(\partial M \times \mathbb{R}_+)$ is a boundary source. Throughout the paper, the coefficients of $\mathcal{A}$, σ, u, f, etc. are real valued. For precise definitions, see section 2.

In this paper we show how to construct Robin boundary values f such that at time T the wave $(u(T), u_t(T))$ is arbitrarily close to $(c\delta_{\widehat{x}}, 0)$, where $\delta_{\widehat{x}}$ is the Dirac delta distribution at a suitable point $\widehat{x} \in M$. We call such waves *focusing waves.* To construct such boundary values, we assume only that we can make physical measurements on a part Γ of the boundary of M. That is, for given Robin values on Γ we can measure the Dirichlet boundary values of the wave on Γ. A focusing wave can then be

[*] Received by the editors October 12, 2007; accepted for publication (in revised form) October 10, 2008; published electronically February 25, 2009. This research has been partially supported by Tekes project MASIT03, EPSRC EP/D065711/1, and the Academy of Finland Center of Excellence programme 213476.

http://www.siam.org/journals/sicon/48-2/70519.html

[†] Institute of Mathematics, Helsinki University of Technology, P.O. Box 1100, Espoo 02015, Finland (fdahl@cc.hut.fi).

[‡] School of Mathematics, The University of Edinburgh, JCMB Mayfield Road, Edinburgh EH9 3JZ, UK (a.kirpichnikova@gmail.com).

[§] Department of Mathematics and Statistics, P.O. Box 68 (Gustaf Hallstromin katu 2b), FI-00014 University of Helsinki, Finland (Matti.Lassas@hut.fi).

generated by an iterative sequence of measurements. In each iteration step we apply time reversal and other simple operators to measured data and compute boundary values for the next iteration step.

The iteration algorithm in this paper is closely related to time reversal methods. Let us, therefore, shortly discuss the underlying idea and the usually used approximations behind these methods. As a simple example, let us consider a domain M in $\mathbb{R}^3$, where we can measure waves and generate sources on the whole boundary of M. Let us first assume that there is a theoretical point source at $\widehat{x} \in M$, and we measure the wave and/or its normal derivative at the boundary of M. Assume further that we record this signal, reverse it in time, and reemitted into M; see [20]. Then one can show (assuming certain approximations hold, see [20, 19, 16]) that the reemitted wave will travel like the original wave but as if time were running backwards. This causes the reemitted wave to focus near $\widehat{x}$.

This principle can also be used for imaging. To find a small scatterer D in a relatively homogeneous domain, one sends a wave into the domain. If the scatterer is small and the single scattering approximation is justified, the scattered wave corresponds to a wave produced by a point source at D. If we record this scattered signal at the boundary, reverse it in time, and reemit it into the domain, it will focus at the scatterer. Furthermore, this focusing has been observed to be quite stable under perturbations of the medium. Thus, if the reemitted wave is simulated (by computational means) in homogeneous media, it will focus at the location of D. In this way a small scatterer can be found using relatively simple computational methods. The above measurement procedure also shows how to approximately focus a wave onto a scatterer. By iterating this measurement procedure, focusing has been observed to improve, and this iterated measurement procedure is called the *iterated time reversal technique*. It has been studied extensively by Fink. There are also various extensions. For example, if the target area contains multiple scatterers, an iteration scheme can be used to focus the wave on any of the scatterers [40].

Besides imaging, time reversal can be used to focus a wave onto a scatterer, say, inside the human body. One application of this is *litotripsy*, where one breaks down a kidney or bladder stone using a focusing ultrasonic wave. Another application is *hyperthermia*, where a cancer is destroyed by an excessive heat dose generated by a focusing wave. Let us point out that, for the wave equation, there are various methods to estimate material parameters in travel time coordinates using boundary measurements. These methods are, however, quite unstable [3, 30]. Therefore, they might not be suitable for hyperthermia, where safety is crucial. An important question is, therefore, how to focus waves in unknown media. This is the topic of the present work.

For reviews and extensions on time reversal, see the seminal papers of Fink [17, 18, 19]. Time reversal methods have been intensively studied in random heterogeneous media where the statistics of the random media are known; see, e.g., [6, 7, 8, 13, 14]. For time reversal in chaotic cavities, see [21]. For related analysis on time reversal methods, see also [4, 5, 27, 22, 35, 39, 41].

Let us describe the key features of the algorithm in this paper. First, to focus a wave onto a point $\widehat{x}$ inside the media, we assume that (i) $\widehat{x}$ is on a normal geodesic from Γ that is distance minimizing and (ii) the travel time coordinates of $\widehat{x}$ are known from the boundary. Let us emphasize that we do not assume the medium in M to be known. However, if the medium (i.e., the coefficient functions of $\mathcal{A}$) is known, the coordinates appearing in condition (ii) can be determined for any given point of M. If the medium is not known explicitly (but only the Robin-to-Dirichlet map is known), then condition

(ii) means that focusing can be done using the same coordinates in which imaging is usually done. Indeed, in imaging algorithms for anisotropic media, the imaging using waves cannot be done in Euclidean coordinates but only in the travel time coordinates (or boundary normal coordinates) related to wave propagation. This is due to the fact that one can deform the coefficients of the equation by pushing those forward with an arbitrary diffeomorphism $\Phi : M \to M$, $\Phi|_{\partial M} = Id$ without changing the measurements at the boundary. Physically speaking, this means that if we have two objects such that the coefficients in the wave equations modeling these objects have the same representation in some coordinate systems, then all boundary measurements for the objects are the same. This invariance of the boundary measurements and the imaging algorithms are considered in detail in [32]. We also note that, in isotropic media, the relationship between travel time coordinates and Euclidian coordinates in M can be written down explicitly in terms of ordinary differential equations [32, Lemma 4.46]. The important feature of the focusing algorithm discussed in this paper is that, as the algorithm does not rely on media parameters obtained from imaging, errors in imaging do not accumulate into errors in focusing.

Second, the algorithm can focus a wave onto an area having no scatterer. This differs from the usual time reversal iteration which can only focus onto a scatterer located inside the media. Third, the algorithm is computationally cheap. In a sense, all computations are done by the media; there is no need to solve the wave equation, cf. [26]. We will assume that the medium is linear, nondispersive, nondissipative, and frequency-independent and depends smoothly on location. However, we do not need any other approximations like the single scattering approximations to prove that the algorithm works.

A limitation of the present algorithm is that we assume self-adjointness of operator $\mathcal{A}$ and time T when the wave focuses is large enough. We also impose the above geometric conditions on $\widehat{x}$.

The present work is a continuation of [12] where a similar iterative scheme was introduced, for which $u(T)$ focuses to a delta distribution, but the time derivative $u_t(T)$ was uncontrolled. The present work can also be seen as a generalization of so-called retrofocusing in control theory, where the aim is to produce boundary sources giving the same final state as a boundary sources sent before in the medium; see [28, 36]. The methodology in this paper arises from boundary control methods used to study inverse problems in hyperbolic equations [9, 10, 11, 31, 32, 33, 34]. In the present work the seminal theorem of unique continuation due to Tataru [44, 46] plays an important role. This result is a generalization of earlier results due to Robbiano [42] and Hörmander [23].

The outline of this work is as follows. In section 2 we introduce notation and review some relevant results from control theory. We also define the boundary operators that are needed in the iteration scheme. In section 3 we describe the main results (Theorems 3.2 and 3.3) and outline their proofs, and in section 4 we prove these results.

2. Definitions and preliminary results. We assume that M is the closure of an open C^∞-smooth bounded set in $\mathbb{R}^m$ $(m \geq 1)$ with nonempty smooth boundary ∂M or an m-dimensional C^∞-smooth compact manifold with boundary. Furthermore, we assume that M is equipped with a C^∞-smooth Riemannian metric $g = \sum_{jk} g_{jk}\, dx^j \otimes dx^k$. Elements of the inverse matrix of g_{ij} are denoted by g^{ij}. Let $\mathrm{d}V_g$ be the smooth measure

$$\mathrm{d}V_g = |g(x)|^{1/2} dx^1 \wedge \cdots \wedge dx^m,$$

where $|g| = \det([g_{jk}])$. Then $L^2(M)$ is defined by the inner product

$$\langle u, v\rangle = \int_M u(x)v(x)\,\mathrm{d}V,$$

where $\mathrm{d}V = \mu \mathrm{d}V_g$ and $\mu \in C^\infty(M)$ is a fixed strictly positive function on M.

In wave equation (1.1), we assume that $\mathcal{A}$ represents the most general formally self-adjoint elliptic partial differential operator of second order with respect to the above inner product [32]. In local coordinates, $\mathcal{A}$ has the form

$$\mathcal{A}v = -\sum_{j,k=1}^{m} \frac{1}{\mu(x)|g(x)|^{1/2}} \frac{\partial}{\partial x^j}\left(\mu(x)|g(x)|^{1/2} g^{jk}(x)\frac{\partial v}{\partial x^k}\right) + q(x)v,$$

where q is a smooth function $q\colon M \to \mathbb{R}$. For example, if $\mu = 1$ and $q = 0$, then $\mathcal{A}$ reduces to the Riemannian Laplace operator. Let us point out that $\mathcal{A}$ represents media that is linear, nondissipative, nondispersive, and frequency-independent.

On the boundary ∂M, operator ∂_ν is defined by

$$\partial_\nu v = \sum_{j=1}^{m} \mu(x)\nu^j \frac{\partial}{\partial x^j} v(x),$$

where $\nu(x) = (\nu^1, \nu^2, \dots, \nu^m)$ is the interior unit normal vector of the boundary satisfying $\sum_{j,k=1}^m g_{jk}\nu^j\nu^k = 1$. To integrate functions on ∂M we use the measure $\mathrm{d}S$ on ∂M induced by $\mathrm{d}V_g$. If $B \subset \partial M \times \mathbb{R}_+$, we define

$$L^2(B) = \{f \in L^2(\partial M \times \mathbb{R}_+) : \operatorname{supp}(f) \subset B\}$$

identifying functions and their zero continuations.

Suppose $\Sigma \subset \partial M$ is a nonempty open set of ∂M, and suppose that $f \in L^2(\Sigma \times \mathbb{R}_+)$. Then wave equation (1.1) has a solution, and we denote this solution by u^f. The map $f \mapsto u^f$ is linear over $\mathbb{R}$, and $\partial_t u^f = u^{\partial_t f}$ when $f, \partial_t f \in L^2(\Sigma, \mathbb{R}_+)$.

The *characteristic function* of a set S is denoted by χ_S.

2.1. Travel time metric. Let $d(x, y)$ be the *geodesic distance* corresponding to g. The metric d is also called the *travel time metric* because it describes how solutions to the wave equation propagate. By the finite velocity of wave propagation (see [25]), we have that if $\Sigma \subset \partial M$ is open and $f \in L^2(\Sigma \times \mathbb{R}_+)$, then at time $t > 0$ solution u^f is supported in the *domain of influence*

$$M(\Sigma, t) = \{x \in M : \mathrm{d}(x, \Sigma) \le t\}.$$

The set $M(x, t)$ is defined by the same formula when $x \in M$. The *diameter* of M is defined as

$$\operatorname{diam}(M) = \max\{\mathrm{d}(x, y) : x, y \in M\}.$$

If $x \in M$ and $\xi \in T_xM$, then we denote by $\gamma_{x,\xi}$ the arc length parametrized geodesic in M that satisfies $\gamma_{x,\xi}(0) = x$ and $\dot\gamma_{x,\xi}(0) = \xi$. Suppose $z \in \Gamma$ for an open set $\Gamma \subset \partial M$ and $\nu = \nu(z)$ is the interior unit normal vector at $z \in \partial M$. Then a geodesic $\gamma_{z,\nu}$ is called a *normal geodesic*, and there is a critical value $\tau_\Gamma(z) > 0$ such that for $t < \tau_\Gamma(z)$ geodesic $\gamma_{z,\nu}([0, t])$ is the unique shortest curve in M that connects $\gamma_{z,\nu}(t)$ to $\overline{\Gamma}$ and for $t > \tau_\Gamma(z)$ this is no longer true. More precisely, we define

$$\tau_\Gamma(z) = \sup\left\{s > 0 : d(\gamma_{z,\nu}(s), \Gamma) = s, \text{ and } \gamma_{z,\nu}(0, s) \subset M^{\mathrm{int}}\right\},$$

where $M^{\rm int}$ are the interior points of M. When $t = \tau_\Gamma(z)$, we say that t is a *critical value* corresponding to $z \in \Gamma$ and $x = \gamma_{z,\nu}(\tau_\Gamma(z))$ is a *cut point* corresponding to Γ. The union of all cut points x is called the *cut locus* with respect to Γ [15].

2.2. Controllability for wave equation. The seminal result implying controllability is Tataru's unique continuation result [44, 46].

PROPOSITION 2.1 (Tataru). *Let $u \in H^1_{\rm loc}(M \times \mathbb{R}_+)$ be a solution of the wave equation*

$$u_{tt}(x,t) + \mathcal{A}u(x,t) = 0.$$

Assume that

$$u|_{\Sigma\times(0,2\tau)} = 0, \quad \partial_\nu u|_{\Sigma\times(0,2\tau)} = 0,$$

where $\Sigma \subset \partial M$ is a nonempty open set and $\tau > 0$. Then

$$u(x,\tau) = 0,\ \partial_t u(x,\tau) = 0 \textit{ for } x \in M(\Sigma,\tau).$$

Using Tataru's unique continuation result, one can prove the following controllability result. The proof is postponed to section 4.

PROPOSITION 2.2 (approximate global controllability). *Suppose Γ is a nonempty open subset of ∂M. If $T > 2\,\mathrm{diam}(M)$, then the linear subspace*

$$\left\{ \left(u^f(T), u^f_t(T) \right) : f \in C_0^\infty(\Gamma \times (0,T)) \right\}$$

is dense in $H^1(M) \times L^2(M)$.

This result yields the following controllability result; see, e.g., [32] and references therein.

PROPOSITION 2.3 (approximative local controllability). *Let $\tau > 0$, let $\Gamma \subset \partial M$ be a nonempty open subset, let $\Gamma_1, \dots, \Gamma_J \subset \Gamma$ be nonempty open sets, and let $0 < s_k < \tau$ for $k = 1, \dots, J$. Suppose*

$$B = \bigcup_{j=1}^{J} \Gamma_j \times (\tau - s_j, \tau)$$

and P is multiplication by the characteristic function χ_B,

$$\begin{aligned} P\colon L^2(\Gamma \times (0,\tau)) &\to L^2(\Gamma \times (0,\tau)), \\ f(x,t) &\mapsto \chi_B(x,t)\, f(x,t). \end{aligned}$$

Then the linear subspace

$$\left\{ u^{Ph}(\tau) : h \in L^2(\Gamma \times (0,\tau)) \right\}$$

is dense in $L^2(N)$, where $N = \bigcup_{j=1}^J M(\Gamma_j, s_j)$.

2.3. Operators for boundary sources. In this section we introduce operators for manipulating boundary sources. These will be needed both in the proof of the main result and in the iteration scheme. Hereafter, we always assume that Γ is a nonempty open set of ∂M where we can control the boundary sources.

For initial boundary value problem (1.1) we define the *nonstationary Robin-to-Dirichlet map* (or *response operator*) Λ by setting

$$\Lambda f = u^f|_{\partial M\times\mathbb{R}_+}, \quad f \in L^2(\partial M \times \mathbb{R}_+).$$

In other words, we solve the wave equation (1.1) for a boundary source f and measure boundary values for the solution u^f. In this work we need only the finite time Robin-to-Dirichlet map restricted onto Γ. For $T > 0$ we define

$$\Lambda^\Gamma_{2T} f = u^f|_{\Gamma\times(0,2T)}, \quad f \in L^2(\Gamma \times (0, 2T)).$$

The map

$$\Lambda^\Gamma_{2T}\colon L^2(\Gamma \times (0, 2T)) \to H^s(\Gamma \times (0, 2T))$$

is bounded, where $H^s(\Gamma \times (0, 2T))$ is the Sobolev space on $\Gamma \times (0, 2T)$ and $s = 1/3$ if the dimension of M is two or larger [45, Theorem 9] and $s < 1/5$ if the dimension is one. To see this, let $\phi \in C^\infty(M)$ be a smooth function with boundary value $f|_{\partial M} = \sigma$, where σ is defined in (1.1). Then $v = e^\phi u^h$ satisfies a wave equation with Neumann boundary condition $\partial_\nu v = 0$, and the mapping properties of Λ^Γ_{2T} follow by [38, Theorem A].

For $f \in L^2(\Gamma \times (0, 2T))$, let

$$\begin{aligned} R_{2T} f(x, t) &= f(x, 2T - t),\\ J_{2T} f(x, t) &= \int_{[0,2T]} J_{2T}(s, t)f(x, s)ds, \end{aligned}$$

where $J_{2T}(s, t) = \frac{1}{2}\chi_L(s, t)$ and

$$L = \{(s, t) \in \mathbb{R}_+ \times \mathbb{R}_+ :\ t + s \le 2T,\ s > t\}.$$

We call R_{2T} the *time reversal map* and J_{2T} the *time filter map* [12]. On $L^2(\Gamma\times(0, 2T))$ with measure $dS(x)dt$, the adjoint of Λ^Γ_{2T} is [12]:

$$\left(\Lambda^\Gamma_{2T}\right)^* = R_{2T}\Lambda^\Gamma_{2T}R_{2T}.$$

The above identity follows since the Green's function $G(x, y, t)$ for problem (1.1) satisfies the reciprocity condition $G(x, y, t) = G(y, x, t)$ when $x, y \in M$ and $t \ge 0$. For $f \in L^2(\Gamma \times (0, 2T))$, let

$$Q_{2T} f = \int_0^{2T} g(t, s)f(x, s)ds \tag{2.1}$$

be the time filter operator, where $g\colon (0, 2T)^2 \to \mathbb{R}$,

$$g(t, s) = \frac{1}{2\left(e^{4T} - 1\right)}\begin{cases}\left(e^t - e^{-t}\right)\left(e^{4T}e^{-s} - e^s\right), & t < s,\\ \left(e^s - e^{-s}\right)\left(e^{4T}e^{-t} - e^t\right), & t > s,\end{cases}$$

is the Green's function for the problem

$$\begin{cases}\left(1 - \partial_t^2\right) g(t, s) = \delta(t - s), & t \in (0, 2T),\\ g|_{t=0} = 0, \quad g|_{t=2T} = 0,\end{cases}$$

where $s \in (0,2T)$. Next, we consider Λ_{2T}^Γ, R_{2T}, J_{2T}, and Q_{2T} as operators:

$$\Lambda_{2T}^\Gamma,\ R_{2T},\ J_{2T},\ Q_{2T} \colon L^2(\Gamma \times (0,2T)) \to L^2(\Gamma \times (0,2T)).$$

Below, we often denote R_{2T}, J_{2T}, and Q_{2T} by R, J, and Q, respectively. For $f, h \in L^2(\Gamma \times (0,2T))$ the *Blagovestchenskii identity* states that

$$\int_M u^f(T) u^h(T)\, dV = \int_{\Gamma \times [0,2T]} (Kf)(x,t) h(x,t)\, dS(x) dt, \tag{2.2}$$

where $K \colon L^2(\Gamma \times (0,2T)) \to L^2(\Gamma \times (0,2T))$ is the bounded operator

$$K = R_{2T}\Lambda_{2T}^\Gamma R_{2T} J_{2T} - J_{2T}\Lambda_{2T}^\Gamma.$$

For a proof, see, e.g., [12] or [32, Lemma 4.15]. The importance of this identity is that it shows how the inner product of solutions $u^f(T)$ and $u^h(T)$ on M can be calculated from data on Γ alone. Namely, on the right-hand side of the Blagovestchenskii identity (2.2), dS is the Riemannian surface volume on ∂M and K is defined in terms of the Robin-to-Dirichlet map Λ_{2T} and simple operators on boundary values like time reversal. Let us point out that since $h \in L^2(\Gamma \times (0,2T))$, the boundary integral is nonzero only on Γ, and, similarly, the left-hand side is nonzero only on $M(\Gamma, T)$. Furthermore, by the Blagovestchenskii identity (2.2), operator K is self-adjoint.

The intrinsic Riemannian surface volume dS on $\Gamma \subset \partial M$ is determined by Λ_{2T}. Namely, by Tataru's unique continuation principle, the Schwartz kernel of Λ_{2T}^Γ is supported in

$$E = \left\{(x,t,x',t') \in (\Gamma \times (0,2T))^2 : t - t' \geq d(x,x')\right\},$$

and the boundary ∂E is in the support. The set ∂E determines the distances of points $z, z' \in \Gamma$ in the same component of Γ with respect to the intrinsic metric of the boundary $(\Gamma, g_{\partial M})$.

3. Iterations and main results.

3.1. Cutoff of wave.

In this section we describe Theorem 3.2 which can be seen as a lemma used in the proof of Theorem 3.3.

For $T > 0$, let X be the Hilbert space

$$X = L^2(\Gamma \times (0,2T)) \times Y, \quad Y = H_0^1\left((0,2T); L^2(\Gamma)\right),$$

with inner product

$$\left\langle \begin{pmatrix} h_1 \\ a_1 \end{pmatrix}, \begin{pmatrix} h_2 \\ a_2 \end{pmatrix} \right\rangle_X = \langle h_1, h_2 \rangle_{L^2} + \langle a_1, a_2 \rangle_{L^2} + \langle \partial_t a_1, \partial_t a_2 \rangle_{L^2}.$$

Here, $H_0^1((0,2T); L^2(\Gamma))$ is the closure of $C_0^\infty((0,2T) \times \Gamma)$ in the Sobolev space $H^1((0,2T); L^2(\partial M))$. As operator Q in (2.1) is the inverse of $1 - \partial_t^2$, it can be considered as an operator $Q \colon Y^* \to Y$, where Y^* is the dual of Hilbert space Y. Furthermore, if $f \in Y$ and $g \in Y^*$, then

$$\langle f, g \rangle_{Y,Y^*} = \langle f, Qg \rangle_Y. \tag{3.1}$$

DEFINITION 3.1. *Let $T > 2\operatorname{diam}(M)$, and let*

$$B = \bigcup_{j=1}^J \Gamma_j \times (T - s_j, T),$$

where $\Gamma_1, \dots, \Gamma_J \subset \Gamma$ are nonempty open sets, and $0 < s_k < T$ for $k = 1, \dots, J$. Let $P = \chi_B$ be the multiplication by the characteristic function of B as in Proposition 2.3, *and let $L\colon X \to X$ be the operator*

$$L = \begin{pmatrix} 1 & 0 \\ 0 & Q \end{pmatrix} \begin{pmatrix} 2PKP & -PK \\ -KP & K - \partial_t K \partial_t \end{pmatrix}. \tag{3.2}$$

Let $\alpha \in (0,1)$, let $\omega > 0$ be such that $2(1 + \|L\|_X) < \omega$, and let

$$S = \left(1 - \frac{\alpha}{\omega}\right) I - \frac{1}{\omega} L.$$

If $f \in L^2(\Gamma \times \mathbb{R}_+)$ is a boundary source, we define a sequence $\binom{h_n}{a_n} = \binom{h_n(\alpha)}{a_n(\alpha)} \in X$, $n = 1, 2, \dots,$ by

$$\begin{cases} \begin{pmatrix} h_0 \\ a_0 \end{pmatrix} = \dfrac{1}{\omega} \begin{pmatrix} PKf \\ 0 \end{pmatrix}, \\ \begin{pmatrix} h_n \\ a_n \end{pmatrix} = \begin{pmatrix} h_0 \\ a_0 \end{pmatrix} + S \begin{pmatrix} h_{n-1} \\ a_{n-1} \end{pmatrix}, \quad n = 1, 2, \dots . \end{cases} \tag{3.3}$$

Theorem 3.2 is the first main result of the present paper. It essentially states that, by a suitable modification of boundary value f, we change the wave so that $u^f(T)$ is multiplied with the characteristic function χ_N for any domain of influence N that can be written as in (3.4) and the time derivative $u_t^f(T)$ is made to vanish. Furthermore, this modification of f relies only on data that can be measured from Γ. This gives an improvement of the retrofocusing technique [28]. For related methods, see [36].

We call iteration (3.3) the *modified time reversal iteration.* The meaning of operator L appearing in it will be seen in (4.4).

Theorem 3.2 (cutoff of wave). *Let $a_1(\alpha), a_2(\alpha), \dots$ be as in Definition* 3.1. *Then the sequence converges in Y,*

$$\lim_{n \to \infty} a_n(\alpha) = a(\alpha),$$

and functions $a(\alpha) \in Y$ on the right-hand side satisfy

$$\lim_{\alpha \to 0} \begin{pmatrix} u^{a(\alpha)}(T) \\ u_t^{a(\alpha)}(T) \end{pmatrix} = \begin{pmatrix} \chi_N u^f(T) \\ 0 \end{pmatrix},$$

where both limits are in $L^2(M)$ and N is the domain of influence

$$N = \bigcup_{k=1}^{J} M(\Gamma_k, s_k). \tag{3.4}$$

Note that here ω may depend on α. For instance, we can choose $\omega = 1/\alpha$.

Let us emphasize that the novelty of this theorem is the explicit iteration scheme for $a(\alpha)$ depending only on boundary measurements. The scheme depends on operators J, Q, P, and K that can be calculated from measurements on Γ. The first three are simple operators like integration and restriction. Operator $K = R_{2T} \Lambda_{2T}^{\Gamma} R_{2T} J_{2T} - J_{2T} \Lambda_{2T}^{\Gamma}$ involves time reversal R_{2T}, time filtering J_{2T}, and two evaluations of the

Robin-to-Dirichlet map Λ_{2T}^{Γ} which correspond to two physical measurements on Γ. Hence, the first order approximation of $a(\alpha)$ requires two physical measurements. After that, each additional pair $(a_n(\alpha), h_n(\alpha))$ requires ten additional measurements. Thus, for approximation $a_n(\alpha)$ we need only finitely many evaluations of the Robin-to-Dirichlet map.

The full proof of Theorem 3.2 is given in section 4.1. Let us here outline the main ideas. For $\alpha \in (0,1)$, boundary sources $h(\alpha), a(\alpha)$ are defined as the minimum of the functional

$$\begin{aligned}\mathcal{F}(h,a,\alpha) = {}& \left\|u^f(T) - u^{Ph}(T)\right\|^2_{L^2(M)} \\ & + \left\|u^{Ph}(T) - u^a(T)\right\|^2_{L^2(M)} + \|u^a_t(T)\|^2_{L^2(M)} \\ & + \alpha \left(\|h\|^2_{L^2(\Gamma\times(0,T))} + \|a\|^2_{L^2(\Gamma\times(0,T))} + \|\partial_t a\|^2_{L^2(\Gamma\times(0,T))}\right). \end{aligned} \tag{3.5}$$

In what follows, when there is no danger of misunderstanding, we denote the L^2-norms in the spaces $L^2(M)$, $L^2(\Gamma\times(0,T))$, etc. just by $\|\cdot\|$. In Lemma 4.1 we use convexity to prove that, for each α, there is a unique minimum $h(\alpha), a(\alpha)$, and by studying the Fréchet derivative of $\mathcal{F}$ we find a linear equation (see (4.2)) for this minimum. In Lemma 4.2 we show that iteration scheme (3.3) converges to a Neumann series that represents the solution to (4.2). That minimizer $a(\alpha)$ satisfies the sought limit is proven in Lemma 4.3. The key step in the proof is to use the approximative controllability results from section 2.2 to show that the first terms in $\mathcal{F}$ can be arbitrarily close to $\|(1-\chi_N)u^f(T)\|^2$ and the next two terms can be made arbitrarily small.

3.2. Focusing of wave. To understand how one can focus waves using Theorem 3.2, suppose we have sets $B \subset \widetilde{B} \subset \Gamma \times (0,T)$ (defined in terms of Γ_i and s_i as in Definition 3.1). Then Theorem 3.2 implies that there are boundary sources $a(\alpha)$ and $\widetilde{a}(\alpha)$ such that

$$\lim_{\alpha\to 0}\begin{pmatrix} u^{a(\alpha)}(T) \\ u_t^{a(\alpha)}(T)\end{pmatrix} = \begin{pmatrix}\chi_N u^f(T) \\ 0\end{pmatrix},$$
$$\lim_{\alpha\to 0}\begin{pmatrix} u^{\widetilde{a}(\alpha)}(T) \\ u_t^{\widetilde{a}(\alpha)}(T)\end{pmatrix} = \begin{pmatrix}\chi_{\widetilde{N}} u^f(T) \\ 0\end{pmatrix},$$

where the domains of influences satisfy $N \subset \widetilde{N}$ at time T. As the solution operator $f \mapsto u^f$ is linear and commutes with ∂_t, solution $b(\alpha) = \widetilde{a}(\alpha) - a(\alpha)$ satisfies

$$\lim_{\alpha\to 0}\begin{pmatrix} u^{b(\alpha)}(T) \\ u_t^{b(\alpha)}(T)\end{pmatrix} = \begin{pmatrix}\chi_{\widetilde{N}\setminus N} u^f(T) \\ 0\end{pmatrix}. \tag{3.6}$$

That is, in the limit, the solution corresponding to b is supported in $\widetilde{N} \setminus N$. In the proof we construct N and $\widetilde{N}$ such that $\widetilde{N} \setminus N$ is a family of sets that shrink onto a chosen point $\widehat{x}$. By further scaling b with a suitable constant depending on the volume of $\widetilde{N} \setminus N$, we obtain the delta distribution.

Notation 1. Let $T > 2\,\mathrm{diam}(M)$ and $\widehat{x} = \gamma_{\widehat{z},\nu}(\widehat{T})$, where $\widehat{z} \in \Gamma$ and $0 < \widehat{T} < T$. Let $\Gamma_j \subset \Gamma$ for $j = 1,2,\ldots$ be open sets around $\widehat{z}$ such that $\Gamma_j \supset \overline{\Gamma}_{j+1}$ and $\bigcap_{j=1}^{\infty} \Gamma_j = \{\widehat{z}\}$.

Suppose $f \in C_0^\infty(\Gamma\times\mathbb{R}_+)$. Let $a_n(\alpha,\varepsilon) \in Y$ be functions obtained from the iteration in Definition 3.1 when B is the set

$$B(\varepsilon) = \Gamma \times \left(T - \left(\widehat{T} - \varepsilon\right), T\right),$$

$\alpha \in (0, \alpha_0)$ for a sufficiently small $\alpha_0 \in (0,1)$, $\varepsilon > 0$, and $\omega = 1/\alpha$. Similarly, let $a_n(\alpha, j, \varepsilon) \in Y$ be functions obtained from the iteration in Definition 3.1 when B is the set

$$B(j,\varepsilon) = \left(\Gamma \times \left(T - \left(\widehat{T} - \varepsilon\right), T\right)\right) \cup \left(\Gamma_j \times \left(T - \widehat{T}, T\right)\right),$$

$\alpha \in (0,1)$, $j = 1, 2, \ldots$, and $\varepsilon > 0$.

Under these assumptions, let

$$b_n(\alpha, j, \varepsilon) = \varepsilon^{-\frac{m+1}{2}} \left(a_n(\alpha, j, \varepsilon) - a_n(\alpha, \varepsilon)\right) \in Y.$$

Theorem 3.3 is the second main result of this paper.

THEOREM 3.3 (focusing wave). *Let $\widehat{z} \in \Gamma$, $\widehat{x} \in M$, $\widehat{T}$, and $b_n(\alpha, j, \varepsilon)$ be as in Notation* 1. *Then functions b_n converge in Y,*

$$\lim_{n \to \infty} b_n(\alpha, j, \varepsilon) = b(\alpha, j, \varepsilon).$$

If $\widehat{T} < \tau_\Gamma(\widehat{z})$, then functions $b(\alpha, j, \varepsilon) \in Y$ satisfy

$$\lim_{\varepsilon \to 0^+} \lim_{j \to \infty} \lim_{\alpha \to 0^+} \begin{pmatrix} u^{b(\alpha,j,\varepsilon)}(T) \\ u_t^{b(\alpha,j,\varepsilon)}(T) \end{pmatrix} = C(\widehat{x}) u^f(T, \widehat{x}) \begin{pmatrix} \delta_{\widehat{x}} \\ 0 \end{pmatrix}, \tag{3.7}$$

where the inner two limits are in $L^2(M)$ and the outer limit is in $\mathscr{D}'(M)$. Furthermore, the constant $C(\widehat{x})$ is nonzero and independent of f, and an explicit expression for $C(\widehat{x})$ is given by (4.10) *in Appendix* A.

If $\Gamma = \partial M$ and $\widehat{T} > \tau_\Gamma(\widehat{z})$, then limit (3.7) *is zero.*

Let us make a few comments. First, we assume that $f \in C_0^\infty(\partial M \times \mathbb{R}_+)$. Hence, $u^f \in C^\infty(M \times \overline{\mathbb{R}}_+)$ (see [37, 38]), and $u^f(\widehat{x}, T)$ exists pointwise. Second, a function $v \in L^2(M)$ is interpreted as a distribution $v \in \mathscr{D}'(M)$ by the formula

$$\langle v, \phi \rangle = \int_M v \phi \, \mathrm{d}V, \quad \phi \in \mathscr{D}(M).$$

Also, the delta distribution at $y \in M$ is defined by $\langle \delta_y, \phi \rangle = \phi(y)$ for $\phi \in \mathscr{D}(M)$.

We say that we can focus a wave onto a point $\widehat{x} = \gamma_{\widehat{z},\nu}(\widehat{T}) \in M$ provided that limit (3.7) in Theorem 3.3 is nonzero. This is the case provided that $\widehat{T} < \tau_\Gamma(\widehat{z})$. Explicitly, we can focus a wave onto any point in the set

$$M_\Gamma := \left\{\gamma_{\widehat{z},\nu}(s) \in M : \widehat{z} \in \Gamma,\, s < \tau_\Gamma(\widehat{z})\right\}.$$

If we have full control of the boundary, that is, if $\Gamma = \partial M$, then we can focus on any point that is not in the cut locus. Let us also point out that the cut locus has zero measure in M [15, 32]. For example, on the closed disc we can focus a wave onto any point except the center. However, if $\Gamma = \partial M$ and $\widehat{T} > \tau_\Gamma(\widehat{z})$, then functions $b_n(\alpha, j, \varepsilon)$ in Theorem 3.3 will be zero for sufficiently large j and small $\varepsilon > 0$ as we will see in the proof. For example, in (3.6) this corresponds to the case when $\widetilde{N} \setminus N$ is empty.

In practice, Theorem 3.3 means that if α, ε are small enough and $\Gamma_j \subset \Gamma$ is small enough, then performing finitely many iterations of the modified time reversal iteration will generate a wave (u, u_t) that at time T is concentrated near a small neighborhood of $\widehat{x}$.

Example. Let us discuss the relation of the modified time reversal iteration and the traditional time reversal. In the traditional time reversal one often considers a wave $v(x,t)$ solving

$$\begin{cases} v_{tt}(x,t) - c(x)^2\Delta v(x,t) = 0 \quad \text{in } \mathbb{R}^m \times \mathbb{R}_+, \\ v|_{t=0} = \phi, \quad v_t|_{t=0} = \psi, \end{cases} \tag{3.8}$$

where ϕ and ψ are supported in a small neighborhood $B(\widehat{x},\rho)$ of $\widehat{x}$. Assume that $M \subset \mathbb{R}^n$ is a bounded domain containing $\widehat{x}$ and that there is T_0 such that $u(x,t) = 0$ for all $x \in M$ and $t > T_0$. This happens, for instance, if dimension m is odd and $c(x) = 1$ in $\mathbb{R}^m$. Let us record the boundary value $h(x,t) := (\partial_\nu - \sigma)v(x,t)|_{\partial M\times(0,T_0)}$ and define the time reversed function $f_0(x,t) = h(x,T_0 - t)$. Then we obtain a boundary value f_0 such that

$$\Big(u^{f_0}(x,T_0), u_t^{f_0}(x,T_0)\Big) = (\phi,\psi), \tag{3.9}$$

where u^{f_0} solves (1.1) with $\mathcal{A} = -c(x)^2\Delta$ and $f = f_0$. As ϕ is supported in a small neighborhood of $\hat{x}$, we can say that the wave focuses near $\hat{x}$ at the time T_0. This focusing of the wave (3.9) at the time T_0 is what makes traditional time reversal interesting. Assume now that $\widehat{x} = \gamma_{\widehat{z},\nu}(\widehat{T})$, where $\widehat{T} < \tau_{\partial M}(\widehat{z})$. If we consider the modified time reversal iteration introduced in Notation 1 with $T = T_0$, $\Gamma_j = \Gamma = \partial M$, $\epsilon > \rho$, and the starting value $f = f_0$, then

$$\lim_{\alpha\to 0}\lim_{n\to\infty} a_n(\alpha,j,\epsilon) = f_1, \quad \lim_{\alpha\to 0}\lim_{n\to\infty} a_n(\alpha,j) = 0.$$

Here, f_1 is a boundary value that produces the same final state (ϕ,ψ) at time T_0 as f_0 but may have a smaller norm in $H^1((0,T_0);L^2(\partial M))$ than f_0. As $\Gamma_j \to \{\widehat{z}\}$ and $\epsilon \to 0$, the modified time reversal iteration produces other waves that focus at time T_0 near $\widehat{x}$ better and better.

4. Proofs.

We start with the proof of Proposition 2.2. The proof is a relatively direct consequence of Tataru's unique continuation theorem and can be found, e.g., in the case of Dirichlet boundary conditions in [36, Lemma 2.1].

Proof of Proposition 2.2. Assume that a pair

$$(\psi,-\phi) \in \big(H^1(M)\times L^2(M)\big)' = H^1(M)'\times L^2(M)$$

satisfies the duality

$$\big\langle u^f(T),\psi\big\rangle_{(H^1(M),H^1(M)')} + \Big\langle u_t^f(T),-\phi\Big\rangle_{L^2(M)} = 0$$

for all $f \in C_0^\infty(\Gamma\times(0,T))$. Note that $H^1(M)$ is the domain of the square root of the operator $\mathcal{A}+cI$ when c is large enough, denoted by $\mathcal{D}(\mathcal{A}^{1/2})$ and $H^1(M)'$ is the dual of $H^1(M) = \mathcal{D}(\mathcal{A}^{-1/2})$. Let

$$\begin{aligned} &e_{tt} + \mathcal{A}e = 0 \quad \text{in } M\times(0,T), \\ &(\partial_\nu - \sigma)e|_{\partial M\times(0,T)} = 0, \quad e|_{t=T} = \phi,\ e_t|_{t=T} = \psi. \end{aligned}$$

By [38, Theorem F], we have that $e \in C([0,T];L^2(M))\cap C^1([0,T];H^1(M)')$ and the trace $e|_{\partial M\times(0,T)} \in H^{-2/5-\epsilon}(\partial M\times(0,T))$, $\epsilon > 0$, is well defined. Thus, we have in sense of distributions

$$\begin{aligned} 0 &= \int_{M\times(0,T)} \left[u^f(e_{tt}+\mathcal{A}e) - \left(u^f_{tt}+\mathcal{A}u^f\right)e\right] \mathrm{d}V\,dt \\ &= \int_M \left(u^f_t(T)\,\phi - u^f(T)\,\psi\right) \mathrm{d}V + \int_{\partial M\times(0,T)} f\,e\,\mathrm{d}S_x\,dt \\ &= \int_{M\times(0,T)} f\,e\,\mathrm{d}S_x\,dt \end{aligned}$$

for all $f \in C_0^\infty(\Gamma\times(0,T))$. This yields that

$$e|_{\Gamma\times(0,T)} = 0 \quad \text{and} \quad \partial_\nu e|_{\partial M\times(0,T)} = 0.$$

To apply unique continuation for $e\in C([0,T];L^2(M))$, let $\epsilon>0$ and let $\eta\in C_0^\infty(\mathbb{R})$ be a function supported on $(-1,1)\subset\mathbb{R}$ whose integral over $\mathbb{R}$ is one. Then

$$e_\epsilon(x,t) = \int_{\mathbb{R}} e(x,t')\eta\left(\frac{t-t'}{\epsilon}\right)dt'$$

satisfies

$$\left(\partial_t^2+\mathcal{A}\right)e_\epsilon = 0 \quad \text{in } M\times(\epsilon,T-\epsilon), \quad (\partial_\nu-\sigma)e_\epsilon|_{\partial M\times(\epsilon,T-\epsilon)} = 0,$$

and $e_\epsilon \in C^\infty((\epsilon,T-\epsilon);L^2(M))$. By representing e_ϵ in terms of eigenfunctions of $\mathcal{A}$ (or by using bootstrap arguments for the wave equation), we see that $e_\epsilon\in C^\infty((\epsilon,T-\epsilon);\mathcal{D}(\mathcal{A}^\infty))\subset C^\infty(M\times(\epsilon,T-\epsilon))$. Now

$$e_\epsilon|_{\Gamma\times(\epsilon,T-\epsilon)} = 0 \quad \text{and} \quad \partial_\nu e_\epsilon|_{\Gamma\times(\epsilon,T-\epsilon)} = 0.$$

Using Tataru's unique continuation theorem [44] we see that if $0<\epsilon<\frac{T}{2}-\text{diam}\,(M)$, then

$$e_\epsilon(x,t) = 0 \quad \text{when } \mathrm{dist}(x,\Gamma) < \frac{T}{2}-\epsilon-\left|t-\frac{T}{2}\right|.$$

In particular, $e_\epsilon(T/2) = \partial_t e_\epsilon(T/2) = 0$. Hence, $e_\epsilon = 0$ identically on $M\times(0,T)$. When $\epsilon\to0$, we see that also e vanishes identically and thus $\phi=\psi=0$. □

4.1. Proof of Theorem 3.2. On X we will study the minimization problem

$$\min_{(h,a)\in X} \mathcal{F}(h,a,\alpha), \tag{4.1}$$

where $\alpha\in(0,1)$ and $\mathcal{F}$ is defined in (3.5). By [38, Theorem A], the map $h\mapsto u^h$ is continuous $L^2(\Gamma\times(0,2T))\to C([0,2T];H^{3/5-\epsilon}(M))$, $\epsilon>0$. Thus, $(h,a)\mapsto\mathcal{F}(h,a,\alpha)$ is a continuous map $X\to\mathbb{R}$.

LEMMA 4.1. *For any $\alpha\in(0,1)$ minimization problem* (4.1) *has a unique minimizer $(h,\,a)\in X$. This minimizer is the unique solution to*

$$(\alpha+L)\begin{pmatrix}h\\a\end{pmatrix} = \begin{pmatrix}PKf\\0\end{pmatrix}, \tag{4.2}$$

where L is defined in (3.2). *Furthermore, $L\colon X\to X$ is nonnegative, bounded, and self-adjoint.*

Proof. By the Blagovestchenskii identity (2.2) we have

$$\begin{aligned}\mathcal{F}(h,a,\alpha) &= \langle f-Ph, K(f-Ph)\rangle + \langle Ph-a, K(Ph-a)\rangle \\ &\quad + \langle \partial_t a, K\partial_t a\rangle \\ &\quad + \alpha(\langle h,h\rangle + \langle a,a\rangle + \langle \partial_t a, \partial_t a\rangle).\end{aligned}$$

Here, K and P are self-adjoint in $L^2(\Gamma \times (0,2T))$. Recall that Y is the Hilbert space $Y = H_0^1((0,2T); L^2(\Gamma))$. By using $\partial_t K \partial_t \colon Y \to Y^*$ and (3.1) it follows that

$$\langle \partial_t a, K\partial_t a\rangle_{L^2(\Gamma\times(0,2T))} = -\langle \partial_t K \partial_t a, a\rangle_{Y^*,Y} = -\langle Q\partial_t K\partial_t a, a\rangle_Y.$$

Thus, using the inner product on X, we can rewrite $\mathcal{F}$ as

$$\begin{aligned}\mathcal{F}(h,a,\alpha) &= \langle f, Kf\rangle + 2\left\langle \begin{pmatrix} h \\ a\end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & Q\end{pmatrix}\begin{pmatrix} -PKf \\ 0\end{pmatrix}\right\rangle_X \\ &\quad + \left\langle \begin{pmatrix} h \\ a\end{pmatrix}, (\alpha + L)\begin{pmatrix} h \\ a\end{pmatrix}\right\rangle_X.\end{aligned} \tag{4.3}$$

As $Q\colon Y^* \to Y$ and $\partial_t K \partial_t \colon Y \to Y^*$ are bounded, $L\colon X \to X$ is bounded. A direct calculation shows that L is self-adjoint. Setting $f = 0$ and $\alpha = 0$ in (4.3) shows that L is nonnegative.

The functional $(h,a) \mapsto \mathcal{F}(h,a,\alpha)$ is strictly convex and satisfies $\mathcal{F}(h,a,\alpha) \geq \frac{\alpha}{2}\|(h,a)\|_X^2 - C_1$ for some $C_1 > 0$. Hence, the unique minimum of $\mathcal{F}$ is at a zero of the Fréchet derivative [2, 47], that is, at $(h,a) \in X$ where $D\mathcal{F}_{h,a} = 0$ and

$$D\mathcal{F}_{h,a}(\xi) = 2\left\langle \begin{pmatrix} 1 & 0 \\ 0 & Q\end{pmatrix}\begin{pmatrix} -PKf \\ 0\end{pmatrix} + (\alpha + L)\begin{pmatrix} h \\ a\end{pmatrix}, \xi\right\rangle, \quad \xi \in X.$$

The Fréchet derivative is invertible as L is nonnegative and self-adjoint, and, thus, $\alpha + L$ is invertible. □

Let us note that if $f = 0$, then (4.3) implies that

$$\begin{aligned}\left\langle \begin{pmatrix} h \\ a\end{pmatrix}, L\begin{pmatrix} h \\ a\end{pmatrix}\right\rangle_X &= \mathcal{F}(h,a,0) \\ &= \left\|u^{Ph}(T)\right\|_{L^2(M)}^2 + \left\|u^{Ph}(T) - u^a(T)\right\|_{L^2(M)}^2 + \|u_t^a(T)\|_{L^2(M)}^2 \\ &\geq \frac{1}{4}\left(\left\|u^{Ph}(T)\right\|_{L^2(M)}^2 + \|u^a(T)\|_{L^2(M)}^2 + \|u_t^a(T)\|_{L^2(M)}^2\right).\end{aligned} \tag{4.4}$$

LEMMA 4.2. *Iteration scheme* (3.3) *converges to the unique solution to* (4.2).

Proof. Using S and ω defined in Definition 3.1, we may rewrite (4.2) as

$$(I - S)\begin{pmatrix} h \\ a\end{pmatrix} = \frac{1}{\omega}\begin{pmatrix} PKf \\ 0\end{pmatrix}.$$

By the nonnegativity of L, it follows that $0 < \langle x, Sx\rangle < 1$ when $\|x\| = 1$. Hence, as S is self-adjoint,

$$\|S\| \leq 1 - \frac{\alpha}{\omega} < 1,$$

and h, a can be written as a convergent Neumann series. □

Lemma 4.3. *Minimizer $h(\alpha), a(\alpha) \in X$ to* (4.1) *satisfy*

$$\begin{aligned} \textit{supp } h(\alpha) &\subset B, \\ a(\alpha) &\in \operatorname{Range}(Q), \\ \lim_{\alpha \to 0} \begin{pmatrix} u^{a(\alpha)}(T) \\ u_t^{a(\alpha)}(T) \end{pmatrix} &= \begin{pmatrix} \chi_N u^f(T) \\ 0 \end{pmatrix}, \end{aligned}$$

where both limits are in $L^2(M)$.

Proof. The first two claims follow by writing out (4.2) with the explicit form of L given in (3.2). For the other results, let us define $Z\colon X \to \mathbb{R}$ by

$$\begin{aligned} Z(h,a) = \frac{1}{2} \left\| \chi_N u^f(T) - u^{Ph}(T) \right\|^2 + \frac{1}{4} \left\| u^a(T) - \chi_N u^f(T) \right\|^2 \\ + \left\| u_t^a(T) \right\|^2 . \end{aligned}$$

To prove the last claim, we show that for any $\varepsilon > 0$ there exists an $\alpha(\varepsilon) \in (0,1)$ such that $Z(h(\alpha), a(\alpha)) < 4\varepsilon$ when $\alpha \in (0, \alpha(\varepsilon))$. By the finite velocity of wave propagation [25],

$$u^{Pf}(T) = \chi_N u^{Pf}(T), \quad f \in L^2(\Gamma \times (0,T)).$$

Hence, for any $(h,a) \in X$,

$$\begin{aligned} \mathcal{F}(Ph, a, \alpha) = {} & \left\| (1-\chi_N) u^f(T) \right\|^2 + \left\| \chi_N u^f(T) - u^{Ph}(T) \right\|^2 \\ & + \left\| u^{Ph}(T) - u^a(T) \right\|^2 + \left\| u_t^a(T) \right\|^2 \\ & + \alpha \left(\|Ph\|^2 + \|a\|^2 + \|\partial_t a\|^2 \right). \end{aligned}$$

It follows that, for any $(h,a) \in X$ and $\alpha \in (0,1)$,

$$Z(Ph, a) \le \mathcal{F}(Ph, a, \alpha) - \left\| (1-\chi_N) u^f(T) \right\|^2 .$$

Here, we have estimated the second term in Z using the triangle inequality and the inequality $(s+t)^2 \le 2(s^2+t^2)$. Let us fix $\varepsilon \in (0,1)$. By Proposition 2.3, there exists an $h_\varepsilon \in L^2(B)$ such that

$$\left\| \chi_N u^f(T) - u^{Ph_\varepsilon}(T) \right\|^2 < \varepsilon,$$

and, by Proposition 2.2, there exists an $a_\varepsilon \in H_0^1((0,2T); L^2(\Gamma))$ such that

$$\begin{aligned} \left\| u^{a_\varepsilon}(T) - \chi_N u^{Ph_\varepsilon}(T) \right\|^2 &< \varepsilon, \\ \left\| u_t^{a_\varepsilon}(T) \right\|^2 &< \varepsilon. \end{aligned}$$

As $h_\varepsilon = Ph_\varepsilon$ we have

$$\mathcal{F}(h_\varepsilon, a_\varepsilon, \alpha) < \left\| (1-\chi_N) u^f(T) \right\|^2 + 3\varepsilon + \alpha \left(\|h_\varepsilon\|^2 + \|a_\varepsilon\|^2 + \|\partial_t a_\varepsilon\|^2 \right),$$

and if $\alpha \in (0, \alpha(\varepsilon))$, where

$$\alpha(\epsilon) = \frac{\varepsilon}{1 + \|h_\varepsilon\|^2 + \|a_\varepsilon\|^2 + \|\partial_t a_\varepsilon\|^2},$$

then minimizer $h(\alpha), a(\alpha$ of $\mathcal{F}$ satisfies

$$\begin{aligned} Z(h(\alpha), a(\alpha)) &\le \mathcal{F}(Ph(\alpha), a(\alpha), \alpha) - \left\| (1-\chi_N) u^f(T) \right\|^2 \\ &\le \mathcal{F}(h_\varepsilon, a_\varepsilon, \alpha) - \left\| (1-\chi_N) u^f(T) \right\|^2 \\ &< 4\varepsilon. \qquad \square \end{aligned}$$

4.2. Proof of Theorem 3.3. By Theorem 3.2, the following limits exist in Y,

$$a(\alpha,\varepsilon) = \lim_{n\to\infty} a_n(\alpha,\varepsilon),$$
$$a(\alpha,j,\varepsilon) = \lim_{n\to\infty} a_n(\alpha,j,\varepsilon)$$

and the following limits exist in $L^2(M)$,

$$\lim_{\alpha\to 0}\begin{pmatrix} u^{a(\alpha,\varepsilon)}(T) \\ u_t^{a(\alpha,\varepsilon)}(T)\end{pmatrix} = \begin{pmatrix}\chi_{N(\varepsilon)}u^f(T) \\ 0\end{pmatrix}, \tag{4.5}$$

$$\lim_{\alpha\to 0}\begin{pmatrix} u^{a(\alpha,j,\varepsilon)}(T) \\ u_t^{a(\alpha,j,\varepsilon)}(T)\end{pmatrix} = \begin{pmatrix}\chi_{N(j,\varepsilon)}u^f(T) \\ 0\end{pmatrix}, \tag{4.6}$$

where

$$N(\varepsilon) = M\left(\Gamma, \widehat{T}-\varepsilon\right),$$
$$N(j,\varepsilon) = M\left(\Gamma, \widehat{T}-\varepsilon\right) \cup M\left(\Gamma_j, \widehat{T}\right).$$

We define $b(\alpha,j,\varepsilon) = \lim_{n\to\infty} b_n(\alpha,j,\varepsilon)$, whence

$$b(\alpha,j,\varepsilon) = \varepsilon^{-\frac{m+1}{2}}\left(a(\alpha,j,\varepsilon) - a(\alpha,\varepsilon)\right).$$

Lemma 4.4. *In $L^2(M)$,*

$$\lim_{j\to\infty}\lim_{\alpha\to 0}\begin{pmatrix} u^{b(\alpha,j,\varepsilon)}(T) \\ u_t^{b(\alpha,j,\varepsilon)}(T)\end{pmatrix} = \varepsilon^{-\frac{m+1}{2}}\begin{pmatrix}\chi_{J(\varepsilon)}u^f(T) \\ 0\end{pmatrix},$$

where

$$J(\varepsilon) = M\left(\widehat{z}, \widehat{T}\right) \setminus M\left(\Gamma, \widehat{T}-\varepsilon\right), \quad \varepsilon > 0.$$

Suppose $r > 0$. For $z \in \partial M$, let

$$\mathcal{B}(z,r) = \{y \in \partial M : d_{\partial M}(z,y) < r\},$$

where $d_{\partial M}$ is the distance on manifold ∂M, and, for $x \in M$, let

$$B(x,r) = \{y \in M : d(x,y) < r\}.$$

The below proof relies on the fact that, for any $\varepsilon > 0$, we have $\Gamma_k \subset \mathcal{B}(\widehat{z},\varepsilon)$ for sufficiently large k.

Proof. Since $a \mapsto u^a$ is linear, it suffices to prove that, pointwise,

$$\lim_{j\to\infty}\chi_{M(\Gamma_j,\widehat{T})\setminus M(\Gamma,\widehat{T}-\varepsilon)}(x) = \chi_{J(\varepsilon)}(x), \quad x \in M. \tag{4.7}$$

This is clear for $x \in J(\varepsilon)$ and $x \notin M(\Gamma, \widehat{T}-\varepsilon)$. If $x \notin M(\widehat{z},\widehat{T})$, we show that $x \notin M(\Gamma_j, \widehat{T})$ for large j. Indeed, if $d(x,\widehat{z}) > \widehat{T}$, then

$$\Gamma_l \subset \mathcal{B}\left(\widehat{z}, \frac{d(x,\widehat{z}) - \widehat{T}}{2}\right)$$

for large l. For $y \in \Gamma_l$, $d(x,y) \geq d(x,\widehat{z}) - d(y,\widehat{z}) > \widehat{T}$, so $d(x,\overline{\Gamma}_{l+1}) > \widehat{T}$. Hence, $d(x,\Gamma_{l+1}) = d(x,\overline{\Gamma}_{l+1}) > \widehat{T}$, and $x \notin M(\Gamma_{l+1},\widehat{T})$ and (4.7) follows. □

The next Lemma shows that $J(\varepsilon)$ are sets that shrink onto $\widehat{x}$ in the case when $\widehat{T} < \tau_\Gamma(\widehat{z})$.

Lemma 4.5 (properties of $J(\delta)$). *Let $\widehat{z} \in \Gamma$ and $\widehat{x} = \gamma_{\widehat{z},\nu}(\widehat{T})$ for $\widehat{T} > 0$. If $\widehat{T} < \tau_\Gamma(\widehat{z})$, then, for any $\varepsilon > 0$, there is a $\delta_0 > 0$ such that for $0 < \delta < \delta_0$,*

$$J(\delta) \subset B(\widehat{x},\varepsilon), \tag{4.8}$$

and $\{\widehat{x}\} \subset J(\delta)$ for all δ so that $\bigcap_{\delta>0} J(\delta) = \{\widehat{x}\}$. If $\Gamma = \partial M$ and $\widehat{T} > \tau_\Gamma(\widehat{z})$, then $J(\delta) = \emptyset$ for δ small enough.

Proof. We need only to prove (4.8) for δ_0 as $J(\delta) \subset J(\delta_0)$ for $\delta < \delta_0$. For a contradiction, suppose that $\varepsilon > 0$ and $x_1, x_2, \dots$ is a sequence such that

$$x_j \in J(1/j), \quad x_j \notin B(\widehat{x},\varepsilon).$$

As M is compact, we can move onto a subsequence and assume that x_j converges to $x \in M \setminus B(\widehat{x},\varepsilon)$. Now $d(x_j,\widehat{z}) \leq \widehat{T}$ and $d(x_j,\Gamma) > \widehat{T} - 1/j$, and $x \mapsto d(x,\Gamma)$ is continuous on M. Hence,

$$d(x,\widehat{z}) \leq \widehat{T}, \quad d(x,\Gamma) \geq \widehat{T}, \tag{4.9}$$

and $\widehat{T} \leq d(x,\Gamma) \leq d(x,\widehat{z}) \leq \widehat{T}$. We have shown that $d(x,\widehat{z}) = d(x,\Gamma) = \widehat{T}$. As M is a compact, there is a distance minimizing curve η from $\widehat{z}$ to x that is parametrized by path length and has length $d(x,\widehat{z})$; see [1]. Hence, η is also a shortest curve from x to Γ. In consequence, $\eta((0,s))$ does not intersect ∂M for some $s > 0$, and $\eta((0,s))$ is a geodesic in M^{int}. Then, by [15, section 3.6], $\eta((0,s))$ is necessarily normal to Γ. (Since $\widehat{z}$ is an interior point of Γ, this can be seen by a shortcut argument.) Thus, η and $\gamma_{\widehat{z},\nu}$ coincide on $[0,s)$. Since $\gamma_{\widehat{z},\nu}((0,\widehat{T}])$ is contained in M^{int}, they coincide on $[0,\widehat{T}]$ and $\widehat{x} = x$. This gives a contradiction with (4.9) and the fact that x is a limit point of $\{x_j\}$. Thus, (4.8) is proven. It is also clear that $\widehat{x} \in J(\delta)$ for all $\delta > 0$.

For the last claim, we show that

$$M\left(\widehat{z},\widehat{T}\right) \subset M(\partial M, T_1)$$

for some $\tau_{\partial M}(\widehat{z}) < T_1 < \widehat{T}$. To see this, let $y \in M \setminus M(\partial M, T_1)$ and $T_1 \in (\tau_{\partial M}(\widehat{z}),\widehat{T})$. Now $d(y,\partial M) > T_1$, and the claim follows if

$$d(y,\partial M) + \delta < d(y,\widehat{z})$$

for sufficiently small $\delta > 0$. For a contradiction, assume that $d(y,\partial M) = d(y,\widehat{z})$. Similar arguments as above show that $\gamma_{\widehat{z},\nu}$ is a shortest geodesic from ∂M to y. This is a contradiction since y is beyond the critical point on $\gamma_{\widehat{z},\nu}$ (see [15, section 3.2]). Hence, $d(y,\partial M) < d(y,\widehat{z})$, and the claim follows. □

Proof of Theorem 3.3. Consider first the case when $\widehat{T} < \tau_\Gamma(\widehat{z})$. By Appendix A, the following limit exists:

$$C(\widehat{x}) = \lim_{\varepsilon \to 0} \frac{\mathrm{Vol}(J(\varepsilon))}{\varepsilon^{\frac{m+1}{2}}}. \tag{4.10}$$

Here, $\mathrm{Vol}(A) = \int_A 1 \, \mathrm{d}V$ when $A \subset M$.

Let us also note that $B(\gamma_{\widehat{z},\nu}(\widehat{T}-\frac{\varepsilon}{2}),\frac{\varepsilon}{2})\subset J(\varepsilon)$, so $\mathrm{Vol}(J(\varepsilon))>0$. Thus, as $u^f(T,\cdot)$ is continuous,

$$\lim_{\varepsilon\to 0}\left\langle \frac{1}{\varepsilon^{\frac{m+1}{2}}}\chi_{J(\varepsilon)}u^f(T),\phi\right\rangle = C(\widehat{x})\lim_{\varepsilon\to 0}\frac{1}{\mathrm{Vol}(J(\varepsilon))}\int_{J(\varepsilon)}u^f(T,x)\phi(x)\mathrm{d}V(x)$$
$$= \left\langle C(\widehat{x})u^f(T,\widehat{x})\delta_{\widehat{x}},\phi\right\rangle, \quad \phi\in\mathscr{D}(M).$$

The result follows by [24, Theorem 2.1.8].

In the case when $\widehat{T}>\tau_\Gamma(\widehat{z})$, $J(\epsilon)=\emptyset$ for ϵ small enough, and, thus, the limits (4.5) and (4.6) are the same. Hence, limit (3.7) is zero. □

4.3. Remark. We have shown that using the modified time reversal algorithm it is possible to focus a wave onto a single point in M at time $t=T$. However, we have not analyzed the behavior of total energy or concentration of energy for $t\in(0,T)$. This question is important for medical applications such as litotripsy.

Appendix A. The limit $C(\widehat{x})$. In this appendix we show that the limit $C(\widehat{x})$ in (4.10) exists. When $m=1$, we have

$$C(\hat{x}) = (\sqrt{g}\mu)(\widehat{x}),$$

so we can assume that $m>1$. Then we can introduce boundary normal coordinates U around $\widehat{x}$. See, e.g., [15] or [32, section 2.1.17]. These are local coordinates (h,y) such that a point $p\in M$ near $\widehat{z}$ is represented by (h,y) provided that $d(p,\partial M)=h$ and $y\in\mathbb{R}^{m-1}$ are local coordinates on ∂M for the unique point $q\in\partial M$ such that $d(q,p)=h$. Let us assume that $\widehat{z}=0$ (identifying points in M with their local representations in U). Then $\gamma_{\widehat{z},\nu}(t)=(t,0,\dots,0)$ for $t<\tau_\Gamma(\widehat{z})$, whence $(\widehat{T},0,\dots,0)\in U$, and we may assume that U has the form

$$U=\{(h,y): h\in[0,D_1),\ |y|<D_2\}$$

for constants $\widehat{T}<D_1<\tau_\Gamma(\widehat{z})$ and $D_2>0$. By $|\cdot|$ and $\langle\cdot,\cdot\rangle$ we denote the Euclidean norm and inner product in $\mathbb{R}^{m-1}$.

The sphere

$$\Sigma=\left\{x\in U: d(x,\widehat{z})=\widehat{T}\right\}$$

can be represented as

$$h(y)=\widehat{T}-\frac{1}{2}\langle y,L\cdot y\rangle+\mathscr{O}\left(|y|^3\right) \tag{A.1}$$

for a symmetric matrix $L\in\mathbb{R}^{(m-1)\times(m-1)}$. Here, we may use the implicit function as $d(\widehat{z},\cdot)$ is smooth outside the cut locus and has a nonzero gradient at $\widehat{x}$. The first order term in (A.1) vanishes by Gauss' lemma (see [43, Proposition 2.3]).

Lemma A.1. *Matrix L is positive definite.*

Proof. In boundary normal coordinates U the metric tensor has the form $g=\mathrm{diag}(1,g_{\alpha\beta})$. Then there are constants $0<c_-<c_+$ such that metrics $g_\pm=\mathrm{diag}(1,c_\pm I)$ satisfy $g_-<g(x)<g_+$ $(x\in U)$ in the sense of positive definite matrices. Explicitly, the $\widehat{T}$-spheres of $g_\pm$ are given by

$$h_\pm(y)=\sqrt{\widehat{T}^2-c_\pm|y|^2}.$$

Near $y = 0$, we may, therefore, express the boundaries of all three $\widehat{T}$-spheres as graphs. For induced metric balls, we have $B_+(\widehat{z}, \widehat{T}) \subset B(\widehat{z}, \widehat{T}) \subset B_-(\widehat{z}, \widehat{T})$, so near $y = 0$ we have

$$h_+(y) \leq h(y) \leq h_-(y).$$

Expanding $h_\pm$ into the Taylor series gives

$$\left\langle y, \left(\frac{c_-}{\widehat{T}} - L\right) \cdot y \right\rangle \leq \mathscr{O}\left(|y|^3\right), \quad \left\langle y, \left(L - \frac{c_+}{\widehat{T}}\right) \cdot y \right\rangle \leq \mathscr{O}\left(|y|^3\right).$$

Hence, $\frac{c_-}{\widehat{T}} - L \leq 0$, $L - \frac{c_+}{\widehat{T}} \leq 0$, and L is positive definite. □

For a symmetric positive definite $(n-1) \times (n-1)$ matrix A, let

$$h_A(y) = \widehat{T} - \frac{1}{2}\langle y, A \cdot y\rangle.$$

For $\varepsilon > 0$ so small that $(0, y) \in U$ when $\widehat{T} - \varepsilon \leq h_A(y)$, we define

$$J(\varepsilon, A) = \left\{(h, y) \in U : \widehat{T} - \varepsilon < h \leq h_A(y)\right\}.$$

The Euclidean volume of this elliptical cap $\mathbb{R}^m$ can be calculated explicitly using coordinate transformation $x \mapsto A^{-1} \cdot x$ and integrating in cylindrical coordinates (see [29, section 9.C]). Hence,

$$\text{(A.2)} \qquad \frac{1}{\sqrt{\det A}} \inf_{x \in J(\varepsilon, A)} f \leq \frac{\operatorname{Vol} J(\varepsilon, A)}{\varepsilon^{\frac{m+1}{2}}} \leq \frac{1}{\sqrt{\det A}} \sup_{x \in J(\varepsilon, A)} f,$$

where, in terms of the gamma function $\Gamma \colon \mathbb{R} \to \mathbb{R}$,

$$f(x) = \frac{2}{m+1} \frac{(2\pi)^{\frac{m-1}{2}}}{\Gamma\left(\frac{m+1}{2}\right)} \left(\mu |g|^{1/2}\right)(x), \quad x \in M.$$

Let us next bound Σ by ellipsoids and bound $\operatorname{Vol} J(\varepsilon)$ in terms of $\operatorname{Vol} J(\varepsilon, A)$ for suitable choices for A. Let $\delta_i \in (0, 1)$ be a sequence such that $\delta_i \to 0$. For each i, let $B_i \subset \partial U$ be an open neighborhood of $y = 0$ such that

$$\text{(A.3)} \qquad h_{(1+\delta_i)L}(y) \leq h(y) \leq h_{(1-\delta_i)L}(y), \quad y \in B_i.$$

Then let $\varepsilon_i > 0$ be a sequence such that sets $J(\varepsilon_i, L), J(\varepsilon_i, (1 \pm \delta_i)L)$ are all defined and $\varepsilon_i \to 0$. By possibly making ε_i smaller we may also assume that $J(\varepsilon_i, L)$, $J(\varepsilon_i, (1 \pm \delta_i)L) \subset p^{-1}(B_i)$ for each i where $p \colon U \to \partial M$, $p(h, y) = (0, y)$, is the projection onto the boundary.

Inequality (A.3) now implies that

$$\operatorname{Vol} J(\varepsilon_i, (1+\delta_i)L) \leq \operatorname{Vol} J(\varepsilon_i) \leq \operatorname{Vol} J(\varepsilon_i, (1-\delta_i)L).$$

By inequality (A.2), we obtain

$$\begin{aligned} \frac{1}{\sqrt{\det(1+\delta_i)L}} \inf_{x \in J(\varepsilon_i, (1+\delta_i)L)} f(x) \leq \frac{\operatorname{Vol} J(\varepsilon_i)}{\varepsilon_i^{\frac{m+1}{2}}} & \\ \leq \frac{1}{\sqrt{\det(1-\delta_i)L}} \sup_{x \in J(\varepsilon_i, (1-\delta_i)L)} f(x), & \end{aligned}$$

and the claim follows by taking $i \to \infty$.

Acknowledgments. We thank the anonymous referees for valuable comments improving the paper.

REFERENCES

[1] R. Alexander and S. Alexander, *Geodesics in Riemannian manifolds-with-boundary*, Indiana Univer. Math. J., 30 (1981), pp. 481–488.

[2] R. Abraham, J.E. Marsden, and T. Ratiu, *Manifolds, Tensor Analysis, and Applications*, Appl. Math. Sci. 75, Springer, New York, 2001.

[3] M. Anderson, A. Katsuda, Y. Kurylev, M. Lassas, and M. Taylor, *Boundary regularity for the Ricci equation, geometric convergence, and Gel'fand's inverse boundary problem,* Invent. Math., 158 (2004), pp. 261–321.

[4] C. Bardos, *A mathematical and deterministic analysis of the time-reversal mirror,* in Inside Out: Inverse Problems and Applications, Math. Sci. Res. Inst. Publ. 47, Cambridge University Press, 2003, pp. 381–400.

[5] C. Bardos and M. Fink, *Mathematical foundations of the time reversal mirror,* Asymptot. Anal., 29 (2002), pp. 157–182.

[6] G. Bal and O. Pinaud, *Time reversal based detection in random media,* Inverse Problems, 21 (2005), pp. 1593–1620.

[7] B. Bal and L. Ryzhik, *Time Reversal for Classical Waves in Random Media,* C. R. Acad. Sci. Paris, Ser. I, 333 (2001), pp. 1041–1046.

[8] G. Bal and L. Ryzhik, *Time reversal and refocusing in random media,* SIAM J. Appl. Math., 63 (2003), pp. 1475–1498.

[9] M. Belishev, *An approach to multidimensional inverse problems for the wave equation,* Dokl. Akad. Nauk, 297 (1987), pp. 524–527 (in Russian).

[10] M. Belishev, *Boundary control in reconstruction of manifolds and metrics (the BC method),* Inverse Problems, 13 (1997), pp. R1–R45.

[11] M. Belishev and Y. Kurylev, *To the reconstruction of a Riemannian manifold via its spectral data (BC-method),* Comm. Partial Differential Equations, 17 (1992), pp. 767–804.

[12] K. Bingham, Y. Kurylev, M. Lassas, and S. Siltanen, *Iterative time-reversal control for inverse problems,* Inverse Prob. Imaging, 2 (2008), pp. 63–81.

[13] L. Borcea, G. Papanicolaou, C. Tsogka, and J. Berryman, *Imaging and time reversal in random media,* Inverse Problems, 18 (2002), pp. 1247–1279.

[14] L. Borcea, G. Papanicolaou, and C. Tsogka, *Theory and applications of time reversal and interferometric imaging,* Inverse Problems, 19 (2003), pp. 5139–5164.

[15] I. Chavel, *Riemannian Geometry. A Modern Introduction,* Cambridge Tracts in Math. 108, Cambridge University Press, Cambridge, 2006.

[16] D. Cassereau and M. Fink, *Time-reversal focusing through a plane interface separating two fluids,* J. Acoust. Soc. Am., 96 (1994), pp. 3145–3154.

[17] M. Fink, *Time reversal mirrors,* J. Phys. D: Appl. Phys., 26 (1993), pp. 1333–1350.

[18] M. Fink, *Time reversal of ultrasonic fields,* IEEE Trans. Ultrasonics, Ferroelectrics, Frequency Control, 39 (1992), pp. 555–566 (Part I), pp. 567–578 (Part II), pp. 579–592 (Part III).

[19] M. Fink, *Time-reversal acoustics in complex environments,* Geophys., 71 (2006), pp. SI151–SI164.

[20] M. Fink, D. Cassereau, A. Derode, C. Prada, P. Roux, M. Tanter, J.-L. Thomas, and F. Wu, *Time-reversed acoustics,* Rep. Prog. Phys., 63 (2000), pp. 1933–1995.

[21] M. Fink and J. De Rosny, *Time reversal experiments in random media and in chaotic cavities,* Nonlinearity, 15 (2002), pp. 1–18.

[22] M. Fink and C. Prada, *Acoustics time-reversal mirrors,* Inverse Problems, 17 (2001), pp. R1–R38.

[23] L. Hörmander, *A uniqueness theorem for second order hyperbolic differential equations,* Comm. Partial Differential Equations, 17 (1992), pp. 699–714.

[24] L. Hörmander, *The Analysis of Linear Partial Differential Operators,* I. Distribution Theory and Fourier Analysis, Springer, Berlin, 1990.

[25] L. Hörmander, *The Analysis of Linear Partial Differential Operators,* IV. Grundlehren Math. Wiss. 275, Springer, New York, 1985.

[26] D. Isaacson, *Distinguishability of conductivities by electric current computed tomography,* IEEE Trans. Med. Imaging, MI-5 (1986), pp. 92–95.

[27] D. Isaacson, M. Cheney, and M. Lassas, *Optimal acoustic measurements,* SIAM J. Appl. Math., 61 (2001), pp. 1628–1647.

[28] B.L.G. Jonsson, M.V. de Hoop, M. Gustafsson, and V.H. Weston, *Retrofocusing of acoustic wave fields by iterated time reversal,* SIAM J. Appl. Math., 64 (2004), pp. 1954–1986.

[29] F. Jones, *Lebesgue Integration on Euclidean Spaces*, Jones and Barlett, Sudbury, MA, 1993.

[30] A. Katsuda, Y. Kurylev, and M. Lassas, *Stability of boundary distance representation and reconstruction of Riemannian manifolds,* Inverse Prob. Imaging, 1 (2007), pp. 135–157.

[31] A. Katchalov and Y. Kurylev, *Multidimensional inverse problem with incomplete boundary spectral data,* Comm. Partial Differential Equations, 23 (1998), pp. 55–95.

[32] A. Katchalov, Y. Kurylev, and M. Lassas, *Inverse Boundary Spectral Problems,* Chapman & Hall/CRC, Boca Raton, FL, 2001.

[33] A. Katchalov, Y. Kurylev, and M. Lassas, *Energy measurements and equivalence of boundary data for inverse problems on non-compact manifolds,* Geometric Methods in Inverse Problems and PDE Control, IMA Vol. Math. Appl. 137, C. Croke, I. Lasiecka, G. Uhlmann, M. Vogelius, eds., Springer, New York, 2004, pp. 183–213.

[34] A. Katchalov, Y. Kurylev, M. Lassas, and N. Mandache, *Equivalence of time-domain inverse problems and boundary spectral problem,* Inverse Problems, 20 (2004), pp. 419–436.

[35] M. Klibanov and A. Timonov, *On the mathematical treatment of time reversal,* Inverse Problems, 19 (2003), pp. 1299–1318.

[36] Y. Kurylev and M. Lassas, *Hyperbolic inverse boundary-value problem and time-continuation of the non-stationary Dirichlet-to-Neumann map,* Proc. Roy. Soc. Edinburgh Sect. A, 132 (2002), pp. 931–949.

[37] I. Lasiecka and R. Triggiani, *Sharp regularity theory for second order hyperbolic equations of Neumann type,* I. L_2 *nonhomogeneous data,* Ann. Mat. Pura Appl., 157 (1990), pp. 285–367.

[38] I. Lasiecka and R. Triggiani, *Regularity theory of hyperbolic equations with nonhomogeneous Neumann boundary conditions,* II. *General boundary data,* J. Differential Equations, 94 (1991), pp. 112–164.

[39] T.D. Mast, A.I. Nachman, and R.C. Waag, *Focusing and imaging using eigenfunctions of the scattering operator,* J. Acoust. Soc. Am., 102 (1997), pp. 715–725.

[40] C. Prada and M. Fink, *Eigenmodes of the time reversal operator: A solution to selective focusing in multiple-target media,* Wave Motion, 20 (1994), pp. 151–163.

[41] C. Prada, J.-L. Thomas, and M. Fink, *The iterative time reversal process: Analysis of the convergence,* J. Acoust. Soc. Am., 97 (1995), pp. 62–71.

[42] L. Robbiano, *Théorème d'unicité adapté au contrôle des solutions des problè hyperboliques,* Comm. Partial Differential Equations, 16 (1991), pp. 789–800.

[43] T. Sakai, *Riemannian Geometry,* American Mathematical Society, Providence, RI, 1992.

[44] D. Tataru, *Unique continuation for solutions to PDEs, between Hörmander's theorem and Holmgren's theorem,* Comm. Partial Differential Equations, 20 (1995), pp. 855–884.

[45] D. Tataru, *On the regularity of boundary traces for the wave equation,* Ann. Sc. Norm. Super. Pisa Cl. Sci., 26 (1998), pp. 185–206.

[46] D. Tataru, *Unique continuation for operators with partially analytic coefficients,* J. Math. Pures Appl., 78 (1999), pp. 505–521.

[47] E. Zeidler, *Nonlinear Functional Analysis and its Applications* III, Variational Methods and Optimization, Springer, New York, 1984.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 2, pp. 859–880

© 2009 Society for Industrial and Applied Mathematics

DECENTRALIZED ADAPTIVE SYNCHRONIZATION OF A STOCHASTIC DISCRETE-TIME MULTIAGENT DYNAMIC MODEL[*]

HONG-BIN MA[†]

Abstract. A decentralized adaptive synchronization problem for a simple yet nontrivial discrete-time stochastic model of network dynamics is investigated, which also illustrates a general framework for a class of adaptive control problems for complex systems with uncertainties. To describe synchronization phenomena in noisy environments, several new definitions of synchronization for stochastic systems are given and applied in our model. In the framework proposed, we prove that in four different cases on local goals, including "deterministic tracking," "center-oriented tracking," "loose tracking," and "tight tracking," under mild conditions on noise sequence and communication limits, the agents in the considered model can achieve global synchronization in sense of mean by using local estimators and controllers based on a least-squares (LS) algorithm. These results show that agents in a complex system disturbed by noise with communication limits can autonomously achieve the global goal of synchronization by using local LS-based adaptive controllers while they are pursuing for their local goals.

Key words. adaptive control, decentralized adaptive synchronization, network dynamics, least-squares algorithm, complex system, discrete-time stochastic model, coupling uncertainties

AMS subject classifications. 93C40, 93C55, 93E24, 93E35

DOI. 10.1137/070685610

1. Introduction.

In this work, we will consider a decentralized adaptive synchronization problem for a particular discrete-time stochastic dynamic network with multiple agents. Our work has mainly three motivations: One comes from the recent research on the capability and limitation of the feedback mechanism [18, 33, 34, 35, 36, 47, 48, 49, 53]; one comes from the decades of studies on traditional adaptive control [2, 10, 16, 23, 26, 27] on a single plant; one comes from the hot studies since the 1980s on complex systems (especially complex networks) [1, 3, 4, 14, 25, 28, 31, 32, 38, 40, 41, 43, 44, 46, 52].

The research on the capability and limitation of the feedback mechanism, initiated by Guo (see Guo's plenary talk [19] in International Congress of Mathematicians 2002 for a brief survey), focuses on revealing the fundamental relationship between the internal uncertainties of a plant and the whole feedback mechanism (the set of all possible feedback control laws), and the kernel problems in this direction are *how much uncertainty can be dealt with by the feedback control* and *what are the limitations of the feedback mechanism*. For example, in the seminal work [47], an uncertain system

$$y_{t+1} = f(y_t) + u_t + w_{t+1} \tag{1.1}$$

with internal uncertainties $f(\cdot) \in \mathcal{F}(L)$ is studied, and it is proved that system (1.1) is stabilizable if $L < \frac{2}{3} + \sqrt{2}$ (here Lipschitz constant L can quantitatively measure the size of $\mathcal{F}(L)$). In previous research on the capability and limitation of the feedback mechanism, only internal uncertainties in one single plant are main uncertainties of interests.

[*] Received by the editors March 19, 2007; accepted for publication (in revised form) October 11, 2008; published electronically February 25, 2009. This work was done in the Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences.
http://www.siam.org/journals/sicon/48-2/68561.html

[†] Temasek Laboratories, National University of Singapore, Singapore 117508 (mathmhb@163.com).

The theory of adaptive control has been developed for decades, and many applications of adaptive control can be found. Traditional adaptive control was mainly developed for linear systems at large, although adaptive control for nonlinear systems has gained more interest in the research community of adaptive control than decades ago. However, most studies on adaptive control are still devoted to dealing with various uncertainties in one single plant, and, hence, strategy of centralized control is still the main concern. Among the seminal work on adaptive control for linear systems, a comprehensive study on discrete-time stochastic adaptive control can be found in [10].

As to the study of complex systems, it is an emerging huge area initiated from physical discoveries on some nonlinear phenomena, especially chaos, fractals, solitons, turbulence, cellular automata, etc. With the development of computer technologies, the focus of studies on complex systems was soon shifted to computer simulations by rule-based generation systems. Several popular books [15, 20, 21, 42] on complexity and complex systems significantly attracted researchers from different disciplines and, hence, pushed the interdisciplinary research on complex systems to a wide range of backgrounds. In the study of complex systems, the so-called *complex adaptive systems* theory [20, 21] plays an important role, which mainly focuses on agent-based modeling and simulations rather than rigorous mathematical analysis.

Motivated by the above issues, we try to consider adaptive control of complex systems, and, due to our limited background, we shall put an emphasis on mathematical study for such problems. Due to the complexity characteristics involved, such as nonlinearity, multihierarchy, and uncertainties, comprehensive theory on adaptive control of complex systems has not come out yet, although few efforts [13, 45] have been devoted in this area. To demonstrate the increasing demand for adaptive control of complex systems, we take a simple example in our practical life—many cars running on a crowded road. In this example, from the point of view of automatic control, the drivers of these cars must control their cars to avoid possible collision while keep their cars running normally. Each driver must take actions on the plant (car) with known or unknown internal parameters (e.g., speed) and try to estimate the speeds or "threats" of those neighboring cars so as to make the car follow a local path. Without considering interactions among cars, driving a car is a typical control problem; however, interactions among cars are inevitable, and, hence, this simple example is in fact a typical adaptive control problem of complex systems, rather than a traditional adaptive control problem, because we cannot design the control laws for all drivers in a centralized approach.

To facilitate mathematical study on adaptive control problems of complex systems, the following simple yet nontrivial theoretical framework is adopted in our theoretical study:

(1) The whole system consists of many dynamical agents, and evolution of each agent can be described by a *dynamic equation*, i.e., state equation (with optional output equation), in the form of a differential equation or difference equation. Different agents may have different structures or parameters.

(2) The evolution of each agent may be influenced by other agents, which means that the dynamic equations of agents are coupled in general. Such influence between agents is usually restricted in local range, and the extent or intensity of reaction can be parameterized.

(3) There exist *information limits* for all of the agents: (a) Each agent knows its internal structure and values of internal parameters; however, it does *not* have access to internal structure or parameters of other agents. (b) Each

agent does *not* know the intensity of influence from others. (c) However, every agent can observe the states of neighbor agents besides its own state.

(4) Under the *information limits* above, each agent may utilize all of the information in hand to estimate the intensity of influence and to design local control to change the state of itself, consequently to influence neighbor agents. Then a basic question can be naturally raised: *Is it possible for all of the agents to achieve a global goal based on the local information and local control?*

The framework above provides a basis for the study of adaptive control of complex networks with uncertainties, and it can be extended further; for example, every agent does not know its internal parameters and it must design its control law based on estimating its internal parameters, which is a main task in traditional adaptive control.

In [37], we have studied a multiagent adaptive control problem within the above framework, which is focused on investigating whether local adaptive controllers based on extended-least-square algorithms can guarantee global closed-loop stability of the whole system, and an affirmative theoretical answer had been given there. In this paper, within the same framework, we will study a problem of decentralized adaptive synchronization for a discrete-time stochastic multiagent dynamical system, and this contribution also illustrates a basic methodology to study the adaptive control problem in the proposed framework. The reason why we choose the adaptive synchronization problem as a starting point is that synchronization, a simple global behavior of agents, is a kind of common and important phenomenon in nature (e.g., chaos synchronization has been found to be useful in secure communication), and, hence, synchronization has been extensively investigated or discussed in the literature (e.g., [41, 46, 52]), especially chaos synchronization [1, 4, 14], delayed neural networks synchronization [8, 22], synchronization in coupled maps [25], synchronization in scale-free or small-world dynamical networks [3, 43, 44], synchronization of complex dynamical networks [28, 31, 32], etc. In recent years, several synchronization-related topics (*coordination*, *rendezvous*, *consensus*, *formation*, etc.) have also become active in the research community [5,6,9,11,24,29,30,39,50]. As for adaptive synchronization, it has received the attention of a few researchers in recent years [7, 12, 51, 54], and the existing work mainly focused on deterministic continuous-time systems, especially chaotic systems, by constructing certain update laws to deal with parametric uncertainties and applying classical Lyapunov stability theory to analyze corresponding closed-loop systems.

The main contributions of this paper are three-fold:

(1) *Framework and methodology.* As an example of theoretical study on adaptive control of complex systems, within the general problem framework stated above, by the methodology of *local analysis–global analysis–local analysis*, we shall give a rigorous study for a decentralized adaptive synchronization problem of a simple multiagent model.

(2) *Concepts and techniques.* To describe synchronization for stochastic discrete-time multiagent systems, we shall propose a series of concepts of *synchronization in sense of mean*, which are not seen in previous studies on deterministic continuous-time systems. Generally speaking, it is not easy to establish *synchronization in sense of mean* since no convenient mathematical tools like Lyapunov stability theory can assert such results directly. To overcome theoretical difficulties, based on Guo's profound results [10, 17] on the least-squares (LS) algorithm, the order estimation techniques and the properties of the LS algorithm are key tools in our analysis.

(3) *Algorithms and results.* Decentralized adaptive synchronization for discrete-time stochastic systems is studied for the first time, based on the frequently used LS estimation algorithm and certainty equivalence principle, and we mathematically established results of decentralized adaptive synchronization in four typical cases.

We shall remark that, due to the existence of random noise in our model, the important concept of equilibrium point (usually denoted by $s(t)$ in previous work) does not exist as in deterministic systems; hence, generally it is *not* possible to design adaptation laws and analyze properties of the overall closed-loop system based on the synchronization errors ($e_i(t) = x_i(t) - s(t)$) as in most existing work [28,31,32,51,54].

The remainder of this paper is organized as follows: In section 2, we will formulate the problem of adaptive synchronization in our framework, and then the main results of this paper are presented in section 3, whose rigorous proofs are given in section 4. Later we illustrate several simulation examples in section 5, and finally we give some concluding remarks in section 6.

2. Problem formulation. In the framework above, as a starting point, we will study a simple stochastic discrete-time dynamic network. In this model, there are N subsystems, and every subsystem represents evolution of an agent. We denote by $x_i(t)$ the state of agent i at time t, and, for simplicity, we assume that linear influences among agents exist in this model. For convenience, we define the concepts of "neighbor" and "neighborhood" as follows: Agent j is a *neighbor* of agent i if agent j has influence on agent i. Let $\mathcal{N}_i$ denote the set of all neighbors of agent i and agent i itself. *Neighborhood* $\mathcal{N}_i$ is a concept describing the communication limits between agent i and others. (Note. Agent i is included in set $\mathcal{N}_i$ just for simplicity, which can also make our model a bit more general.)

Model of the system. Suppose that each agent i ($i = 1, 2, \ldots, N$) has the following state equation:

$$x_i(t+1) = f_i(x_i(t)) + u_i(t) + \gamma_i \bar{x}_i(t) + w(t+1), \tag{2.1}$$

where $f_i(\cdot)$ represents the internal structure of agent i, $u_i(t)$ is the local control of agent i, $\gamma_i \bar{x}_i(t)$ reflects the influence of the other agents towards agent i, and $\{w(t)\}$ is the unobservable random noise sequence. Here γ_i denotes the intensity of influence, and $\bar{x}_i(t)$ is the weighted average of states of agents in the neighborhood of agent i, i.e.,

$$\bar{x}_i(t) = \sum_{j \in \mathcal{N}_i} g_{ij} x_j(t), \tag{2.2}$$

where the nonnegative constants $\{g_{ij}\}$ satisfy

$$\sum_{j \in \mathcal{N}_i} g_{ij} = 1. \tag{2.3}$$

In the framework above, agent i does not know the intensity of influence γ_i; however, it can use the historical information

$$\{x_i(t), \bar{x}_i(t), u_i(t-1), x_i(t-1), \bar{x}_i(t-1), u_i(t-2), \ldots, x_i(1), \bar{x}_i(1), u_i(0)\} \tag{2.4}$$

to estimate γ_i and can further try to design its local control $u_i(t)$ to achieve its local goal.

Remark 2.1. As mentioned in the introduction, model (2.1) is partially motivated by recent research on the capability and limitation of the feedback mechanism, and here we want to explore the capability of adaptive control in dealing with coupling uncertainties within multiple subsystems rather than internal uncertainties within one single system. Although model (2.1) is simple enough, it can capture all essential features that we want, and the simple model can be viewed as a prototype or approximation of more complex models. Model (2.1) highlights the difficulties in dealing with coupling uncertainties by feedback control. The ideas in this paper can be also applied in more general or complex models, which may be considered in our future work and may involve more difficulties in the design and theoretical analysis of local adaptive controllers.

Estimation algorithm. In this paper, we assume that each agent is smart enough and it can use the LS algorithm to estimate the unknown intensity of influence. Since the LS algorithm is widely used in statistics, system identification, and adaptive control, we choose the LS algorithm as a starting point to study adaptive control of complex systems. For agent i, we denote by $\hat{\gamma}_i(t)$ the estimated value of γ_i. Denote

$$y_i(t) = x_i(t) - f_i(x_i(t-1)) - u_i(t-1) \tag{2.5}$$

and

$$\begin{aligned} Y_i(t) &= (y_i(1), y_i(2), \ldots, y_i(t))^\tau, \\ \bar{X}_i(t) &= (\bar{x}_i(0), \bar{x}_i(1), \ldots, \bar{x}_i(t-1))^\tau, \\ W(t) &= (w(1), w(2), \ldots, w(t))^\tau; \end{aligned} \tag{2.6}$$

then we have $Y_i(t) = \gamma_i \bar{X}_i(t) + W(t)$. Naturally, define

$$\hat{\gamma}_i(t) = \arg\min_{\gamma} \left\| Y_i(t) - \gamma \bar{X}_i(t) \right\|, \tag{2.7}$$

where (and hereafter) $\|\cdot\|$ represents the Euclidian norm, i.e., $\|v\| = \sqrt{v^\tau v}$. Then it is easy to obtain that

$$\begin{aligned} \hat{\gamma}_i(t) &= \left[\bar{X}_i^\tau(t)\bar{X}_i(t)\right]^{-1}\left[\bar{X}_i^\tau(t)Y_i(t)\right] \\ &= \left[\sum_{k=0}^{t-1} \bar{x}_i^2(k)\right]^{-1}\left[\sum_{k=1}^{t} \bar{x}_i(k-1)y_i(k)\right] \end{aligned} \tag{2.8}$$

which can be transformed into the recursive form

$$\begin{aligned} \hat{\gamma}_i(t+1) &= \hat{\gamma}_i(t) + \bar{a}_i(t)\bar{p}_i(t)\bar{x}_i(t)[y_i(t+1) - \hat{\gamma}_i(t)\bar{x}_i(t)], \\ \bar{p}_i(t+1) &= \bar{p}_i(t) - \bar{a}_i(t)[\bar{p}_i(t)\bar{x}_i(t)]^2 \end{aligned} \tag{2.9}$$

by defining

$$\begin{aligned} \bar{a}_i(t) &\triangleq \left[1 + \bar{p}_i(t)\bar{x}_i^2(t)\right]^{-1}, \\ \bar{p}_i(t) &\triangleq \left[\sum_{k=0}^{t-1} \bar{x}_i^2(k)\right]^{-1}. \end{aligned} \tag{2.10}$$

The recursive LS algorithm (2.9) can efficiently update the parameter estimate $\hat{\gamma}_i(t)$ online without much computation cost. In practical use, the initial values $\hat{\gamma}_i(0)$ can be taken arbitrarily and $0 < \bar{p}_i(0) < \frac{1}{e}$ such that $\bar{p}_i^{-1}(t+1) \geq \bar{p}_i^{-1}(0) > e$. (Hereafter, e is the base of natural logarithm.)

Local goals and local controllers. Due to the limitation in the communication among the agents, generally speaking, agents can only try to achieve local goals. Naturally, we assume that agent i tries to track a local signal $\{z_i(t)\}$, which can be a known sequence or a stochastic sequence relating to other agents. Later we will discuss several different cases. Supposing that agent i knows the intensity of influence from others, i.e., γ_i, in order to track its local goal, local controller of agent i can be naturally given by

$$\hat{u}_i(t) = \arg\min_{u_i(t)} E[x_i(t+1) - z_i(t)]^2 \tag{2.11}$$

which yields

$$\hat{u}_i(t) = -f_i(x_i(t)) - \gamma_i \bar{x}_i(t) + z_i(t) \tag{2.12}$$

by (2.1). Within our framework, agent i knows the function $f_i(\cdot)$ but does *not* know γ_i. Hence, by using the certainty equivalence principle, agent i can use the following adaptive control law:

$$u_i(t) = -f_i(x_i(t)) - \hat{\gamma}_i(t)\bar{x}_i(t) + z_i(t), \tag{2.13}$$

where $\hat{\gamma}_i(t)$ is updated online by recursive LS algorithm (2.9).

Synchronization problem. With the local LS-based adaptive controllers designed via local tracking goals, we want to know *whether all of the agents can autonomously achieve the global goal of synchronization in some sense.* Intuitively, synchronization can be interpreted as follows: For every pair of agents i and j ($i \neq j$), the difference $e_{ij}(t) \overset{\Delta}{=} x_i(t) - x_j(t)$ approaches zero (or its minimum) asymptotically. In our model, due to the presence of random noise, generally $\lim_{t\to\infty} e_{ij}(t) = 0$ cannot be expected. Thus, it is necessary to introduce new concepts of synchronization in sense of mean, some of which are defined in the following.

Definition 2.1. *If the system satisfies*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} |e_{ij}(t)| = 0 \ \forall i \neq j, \tag{2.14}$$

then we say it achieves (strong) synchronization in sense of mean.

Definition 2.2. *If the system satisfies*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} e_{ij}(t) = 0 \ \forall i \neq j, \tag{2.15}$$

then we say it achieves weak synchronization in sense of mean.

Definition 2.3. *If the system satisfies*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} |e_{ij}(t)|^p = 0 \ \forall i \neq j, \tag{2.16}$$

then we say it achieves synchronization in sense of pth mean. Especially when $p = 2$, we will say it achieves synchronization in sense of mean squares.

Remark 2.2. Previous research on network synchronization mainly deals with noise-free systems (see, e.g., [31, 32, 43, 46]), where a special solution (equilibrium point) $s(t)$ can be defined by the dynamics of each agent and, hence, synchronization means $e_i(t) \overset{\Delta}{=} x_i(t) - s(t) \to 0$ as $t \to \infty$. However, in our model, due to the existence of noise and complex control law, such equilibrium point $s(t)$ cannot be well-defined. This is why we introduce notions of $e_{ij}(t)$, based on which synchronization concepts can be defined in both deterministic cases and stochastic cases.

Remark 2.3. We can easily prove that the definitions above have the following connections: *synchronization in sense of pth mean* ($p > 2$) $\Longrightarrow$ *synchronization in sense of mean squares* $\Longrightarrow$ *(strong) synchronization in sense of mean* $\Longrightarrow$ *weak synchronization in sense of mean.*

Remark 2.4. In model (2.1), the noise sequence $\{w(t+1)\}$ is common for all agents. However, more general cases ($w(t+1)$ replaced with $w_i(t+1)$) can be considered without difficulties. For simplicity, we consider only common noise disturbance, which intuitively means that the environment acts on the agents in the same way; this can also make the definitions of synchronization in sense of mean more simple and clear.

Local tracking goals. For the tracking signals $\{z_i(t)\}$, we discuss the following cases in this paper.

Case I (deterministic tracking). $z_i(t) = z^*(t)$, where $\{z^*(t)\}$ is a sequence of deterministic signals (bounded or even unbounded) which satisfies $|z^*(t)| = O(t^\delta)$.

Case II (center-oriented tracking). $z_i(t) = \bar{z}(t)$, where $\bar{z}(t) = \frac{1}{N}\sum_{i=1}^{N} x_i(t)$ is the center state of all agents, i.e., average of states of all agents.

Case III (loose tracking). $z_i(t) = \lambda\bar{x}_i(t)$, where constant $|\lambda| < 1$. This case means that the tracking signal $z_i(t)$ is close to the average of states of neighbor agents of agent i, and factor λ describes how close it is.

Case IV (tight tracking). $z_i(t) = \bar{x}_i(t)$. This case means that the tracking signal $z_i(t)$ is exactly the average of states of agents in the neighborhood of agent i.

In the first two cases, all agents track a common signal sequence, and the only differences are as follows: In Case I the sequence has nothing with every agent's state; however, in Case II the sequence is the center state of all of the agents. The first two cases mean that a common "leader" of all of agents exists, who can communicate with and send commands to all agents; however, the agents can only communicate with one another under certain *information limits.* In Cases III and IV, no common "leader" exists and all agents attempt to track the average state $\bar{x}_i(t)$ of its neighbors, and the difference between them is just the factor of tracking tightness.

In the remainder of this paper, we will consider the decentralized adaptive synchronization problem formulated above in Cases I–IV.

3. Main results. In the above cases, under mild conditions on noise, we shall prove that all agents can achieve synchronization in sense of mean by using the LS-based learning and control algorithm defined above, which demonstrates that agents in a complex system disturbed by noise with "information limits" can exhibit the collective behavior, synchronization, by a properly designed local learning algorithm and local adaptive controllers based on local goals.

In order to analyze the above adaptive synchronization problem, we introduce the following assumption on the noise sequence, which allows for a wide class of stochastic noise.

Assumption A1. The noise sequence $\{w(t), \mathcal{F}_t\}$ is a martingale difference sequence (with $\{\mathcal{F}_t\}$ being a sequence of nondecreasing σ-algebras) such that

$$\sup_t E\left[|w(t+1)|^\beta|\mathcal{F}_t\right] < \infty \quad a.s. \tag{3.1}$$

for a constant $\beta > 2$.

Theorem 3.1. *In Cases* I, II, *and* III, *suppose that system* (2.1) *satisfies Assumption* A1*; then the decentralized LS-based adaptive controller has the following closed-loop properties:*

(1) *All of the agents can asymptotically correctly estimate the intensity of influence from others, i.e.,*

$$\lim_{t\to\infty} \hat{\gamma}_i(t) = \gamma_i. \tag{3.2}$$

(2) *The system can achieve synchronization in sense of mean, i.e.,*

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} |e_{ij}(t)| = 0 \ \forall i \neq j. \tag{3.3}$$

(3) *The system can achieve synchronization in sense of mean squares, i.e.,*

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} |e_{ij}(t)|^2 = 0 \ \forall i \neq j. \tag{3.4}$$

For the synchronization in Case IV, the following assumption is necessary.

Assumption A2. Matrix $G = (g_{ij})$ $[g_{ij} = 0$ if $j \notin \mathcal{N}_i]$ is an irreducible primitive matrix.

Remark 3.1. Assumption A2 excludes those cases that matrix G is reducible. This assumption means that all of the agents should be connected so that they can synchronize with each other in Case IV. The primitiveness of matrix G excludes those cases where matrix G is cyclic (or periodic from the point of view of Markov chain), which should also be avoided for the goal of synchronization in Case IV.

Theorem 3.2. *In Case* IV, *suppose that Assumption* A1 *holds for system* (2.1) *and Assumption* A2 *holds for matrix* $G = (g_{ij})$ *then the decentralized LS-based adaptive controller has the following closed-loop properties:*

(1) *All of the agents can asymptotically correctly estimate the intensity of influence from others, i.e.,*

$$\lim_{t\to\infty} \hat{\gamma}_i(t) = \gamma_i. \tag{3.5}$$

(2) *The system can achieve synchronization in sense of mean, i.e.,*

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} |e_{ij}(t)| = 0 \ \forall i \neq j. \tag{3.6}$$

(3) *The system can achieve synchronization in sense of mean squares, i.e.,*

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} |e_{ij}(t)|^2 = 0 \ \forall i \neq j. \tag{3.7}$$

Remark 3.2. By Remark 3.1, Assumption A2 cannot be weakened in general for the synchronization of all agents in Case IV. From the proof of Theorem 3.2, we can see that it is also the necessary and sufficient condition for the trivial case where there is no noise disturbance and the parameter γ_i is known by agent i. In fact, in this case it is unnecessary to estimate γ_i or to deal with the external disturbance, and, hence, the control law can be taken as $u_i(t) = -f_i(x_i(t)) - \gamma_i \bar{x}_i(t) + \bar{x}_i(t)$ and the closed-loop system is $x_i(t+1) = \bar{x}_i(t), i = 1, 2, \ldots, N$, whose matrix form is $X(t+1) = GX(t)$. Then obviously we cannot guarantee that all elements of $X(t)$ synchronize if G is reducible or cyclic.

4. Proofs of main results.

To make rigorous analysis, for $i = 1, 2, \ldots, N$, denote $\bar{r}_i(t) = 1 + \sum_{k=1}^{t} \bar{x}_i^2(k)$ and $\tilde{\gamma}_i(t) = \gamma_i - \hat{\gamma}_i(t)$. By the recursive LS algorithm, obviously we have $\bar{r}_i(t) = \bar{p}_i^{-1}(t+1) + c_0$, and constant c_0 is determined by $\bar{p}_i^{-1}(0)$.

4.1. Auxiliary lemmas.

We start with several lemmas, which will be used in the proofs of theorems.

LEMMA 4.1. *Under Assumption* A1, *the LS estimator defined by* (2.9) *has the following properties almost surely:*

(1)

$$\bar{p}_i^{-1}(t+1)\tilde{\gamma}_i^2(t+1) + (1+o(1))\sum_{k=1}^{t} \bar{a}_i(k)\left[\tilde{\gamma}_i(k)\bar{x}_i^2(k)\right]^2 = \sigma^2 \sum_{k=1}^{t} \bar{a}_i(k)\bar{p}_i(k)\bar{x}_i^2(k) + o(\log \bar{r}_i(t)) + O(1). \tag{4.1}$$

(2)

$$\bar{p}_i^{-1}(t+1)\tilde{\gamma}_i^2(t+1) = O(\log \bar{r}_i(t)). \tag{4.2}$$

(3)

$$\sum_{k=1}^{t} \bar{a}_i(k)\left[\tilde{\gamma}_i(k)\bar{x}_i(k)\right]^2 = \sum_{k=1}^{t} \frac{\left[\tilde{\gamma}_i(k)\bar{x}_i(k)\right]^2}{1+\bar{p}_i(k)\bar{x}_i^2(k)} = O(\log \bar{r}_i(t)). \tag{4.3}$$

(4) *If*

$$\bar{p}_i(t)\bar{x}_i^2(t) \to 0, \;\; \bar{p}_i^{-1}(t) \to \infty \tag{4.4}$$

as $t \to \infty$, *then*

$$\bar{p}_i^{-1}(t+1)\tilde{\gamma}_i^2(t+1) + \sum_{k=1}^{t}[\tilde{\gamma}_i(k)\bar{x}_i(k)]^2 \sim \sigma^2 \log \bar{r}_i(t). \tag{4.5}$$

Proof. Denote

$$\mathcal{F}_t = \sigma\{w(0), w(1), \ldots, w(t)\}. \tag{4.6}$$

By (2.1), (2.13), and (2.9), obviously $x_i(t), \bar{x}_i(t) \in \mathcal{F}_t$. Hence, the properties of the LS algorithm (see [10,17]) can be applied. This lemma is just the special one-dimensional case. Equation (4.1) corresponds to [17, Theorem 6.3.1], and (4.2)–(4.4) correspond to Corollaries 6.3.1, 6.3.2, and 6.3.3 of [17]. □

By (4.2), immediately we have the following corollary.

COROLLARY 4.1. *The estimation* $\hat{\gamma}_i(t)$ *of* γ_i *converges to the true value* γ_i *almost surely with the convergence rate*

$$|\tilde{\gamma}_i(t)| = O\left(\sqrt{\tfrac{\log \bar{r}_i(t)}{\bar{r}_i(t)}}\right). \tag{4.7}$$

LEMMA 4.2. *Under Assumption* A1, *for* $i = 1, 2, \ldots, N$, *we have*

$$\sum_{k=1}^{t}[x_i(k)]^2 \to \infty, \;\; \bar{r}_i(t) \to \infty, \;\; \bar{p}_i(t) \to 0 \textit{ as } t \to \infty \textit{ a.s.} \tag{4.8}$$

Proof. Under assumption A1, this lemma can be established by applying the martingale estimation theorem to the terms in

$$\sum_{k=1}^{t}[x_i(k+1)]^2 = \sum_{k=1}^{t}[g_i(k)]^2 + \sum_{k=1}^{t}[w(k+1)]^2 + 2\sum_{k=1}^{t} g_i(k)w(k+1). \tag{4.9}$$

Details are omitted here to save space. □

Lemma 4.3. *Assume the nonnegative sequences $\{X_t\}$ and $\{d_t\}$ satisfy $X_{t+1} = O(\max(X_t, d_t))$ as $t \to \infty$. Denote*

$$S_t = \sum_{k=1}^{t} |X_k|, \quad D_t = \sum_{k=1}^{t} d_k. \tag{4.10}$$

If $S_t \to \infty$ as $t \to \infty$, we can get

$$S_{t+1} = O(S_t + D_t), \tag{4.11}$$

and the "O" constant satisfies

$$\limsup_{t\to\infty} \frac{S_{t+1}}{S_t + D_t} \le \limsup_{t\to\infty} \frac{X_{t+1}}{X_t + d_t}. \tag{4.12}$$

In addition, if $d_{t+1} = O(d_t)$ as $t \to \infty$, then

$$D_{t+1} = O(D_t), \quad S_{t+1} + D_{t+1} = O(S_t + D_t), \tag{4.13}$$

and

$$\limsup_{t\to\infty} \frac{S_{t+1} + D_{t+1}}{S_t + D_t} \le \limsup_{t\to\infty} \frac{X_{t+1} + d_{t+1}}{X_t + d_t}. \tag{4.14}$$

Proof. According to definitions of the notations $O(\cdot)$ and $o(\cdot)$, this lemma can be proved by using $\epsilon - \delta$ language without much difficulty, and, hence, the details are omitted here to save space. □

Lemma 4.4. *Consider the following iterative system:*

$$X_{t+1} = A_t X_t + W_t, \tag{4.15}$$

where $\|W_t\| = O(t^\delta)$, δ is an arbitrary nonnegative constant, and $A_t \to A$ as $t \to \infty$. Assume ρ is the spectral radius of A, i.e., $\rho = \max\{|\lambda(A)|\}$. Denote $S_t = 1 + \sum_{k=1}^{t} ||X_k||^2$. For arbitrary $\epsilon > 0$,

$$\begin{aligned} \|X_t\| &= o\left(t^\delta(\rho+\epsilon)^t\right) + O\left(t^\delta\right), \\ S_t &= o\left(t^{2\delta+1}(\rho+\epsilon)^{2t}\right) + O\left(t^{2\delta+1}\right). \end{aligned} \tag{4.16}$$

Furthermore,

(1) *if $\rho \in [0, 1)$, we can get*

$$\frac{\log S_t}{t} = O\left(\frac{\log t}{t}\right) = o(1); \tag{4.17}$$

(2) *if $\rho = 1$, we can get*

$$\frac{\log S_t}{t} = o(1); \tag{4.18}$$

(3) *if* $\rho > 1$, *we can get*

$$\frac{\log S_t}{t} = O(1). \tag{4.19}$$

Proof. For arbitrary $\epsilon > 0$, by the definition of ρ, there exists a matrix norm (denoted by $\|\cdot\|_p$) such that $\|A\|_p < \rho + \frac{\epsilon}{2}$; we can get also $\|A_t\|_p \to \|A\|_p$ from $A_t \to A$ as $t \to \infty$. Hence, for sufficiently large t,

$$\|A_t\|_p < \|A\|_p + \frac{\epsilon}{2} < \rho + \epsilon. \tag{4.20}$$

According to the equivalence among norms, $\|W_t\|_p = O(\|W_t\|) = O(t^\delta)$; therefore, for sufficiently large t,

$$\|X_{t+1}\|_p \le \|A_t\|_p\|X_t\|_p + \|W_t\|_p \le (\rho+\epsilon)\|X_t\|_p + C_p t^\delta. \tag{4.21}$$

Iterating the inequality above, we have

$$\begin{aligned}\|X_t\|_p &\le C_p \sum_{k=1}^{t-m} (\rho+\epsilon)^{k-1}(t-k)^\delta + (\rho+\epsilon)^{t-m}\|X_m\|_p \\ &\le C_p t^\delta \sum_{k=1}^{t-m} (\rho+\epsilon)^{k-1} + (\rho+\epsilon)^{t-m}\|X_m\|_p,\end{aligned} \tag{4.22}$$

where m is a constant depending on ϵ and p. Obviously

$$\begin{aligned}\|X_t\|_p &= O\left(t^\delta\left[(\rho+\epsilon)^t + O(1)\right]\right) + O\left((\rho+\epsilon)^t\right) \\ &= O\left(t^\delta(\rho+\epsilon)^t + t^\delta\right).\end{aligned} \tag{4.23}$$

By the arbitrariness of ϵ and the equivalence among norms, we can get

$$\|X_t\| = o\left(t^\delta(\rho+\epsilon)^t\right) + O\left(t^\delta\right). \tag{4.24}$$

By the definition of S_t and the equivalence among norms, we have

$$\begin{aligned}S_t &= O\left(\sum_{k=1}^{t} \|X_t\|_p^2\right) \\ &= O\left(\sum_{k=1}^{t}\left[k^{2\delta}(\rho+\epsilon)^{2k} + k^{2\delta}\right]\right) \\ &= O\left(t^{2\delta+1}\left[(\rho+\epsilon)^{2t} + O(1)\right] + t^{2\delta+1}\right) \\ &= O\left(t^{2\delta+1}(\rho+\epsilon)^{2t} + t^{2\delta+1}\right).\end{aligned} \tag{4.25}$$

Furthermore, by the arbitrariness of ϵ,

$$S_t = o\left(t^{2\delta+1}(\rho+\epsilon)^{2t}\right) + O\left(t^{2\delta+1}\right). \tag{4.26}$$

Consequently, validity of this lemma can be easily established in all cases. □

Lemma 4.5. *Consider the following iterative system:*

$$X_{t+1} = A_t X_t + W_t, \tag{4.27}$$

where $A_t \to A$ as $t \to \infty$ and $\{W_t\}$ satisfies

$$\sum_{k=1}^{t} \|W_k\|^2 = o(t). \tag{4.28}$$

If the spectral radius $\rho(A) < 1$, then

$$\sum_{k=1}^{t} \|X_k\| = o(t), \ \ \sum_{k=1}^{t} \|X_k\|^2 = o(t). \tag{4.29}$$

Proof. By Schwartz's inequality,

$$\sum_{k=1}^{t} \|W_k\| \le \left(\sum_{k=1}^{t} \|W_k\|^2\right)^{\frac{1}{2}} \left(\sum_{k=1}^{t} 1^2\right)^{\frac{1}{2}} = o(t). \tag{4.30}$$

Since $\rho < 1$, we can take a small number $\epsilon > 0$ such that $\rho + \epsilon < 1$, and there exists a matrix norm $\|\cdot\|_p$ such that $\|A\|_p < \rho + \frac{\epsilon}{2}$; in addition, $\|A_t\|_p \to \|A\|_p$ because of $A_t \to A$ as $t \to \infty$, and, hence, for sufficiently large t,

$$\|A_t\|_p < \|A\|_p + \frac{\epsilon}{2} < \rho + \epsilon. \tag{4.31}$$

By the equivalence among norms, $\|W_t\|_p = O(\|W_t\|)$, and, hence, for sufficiently large t, we have

$$\|X_{t+1}\|_p \le \|A_t\|_p \|X_t\|_p + \|W_t\|_p \le (\rho + \epsilon)\|X_t\|_p + \|W_t\|_p. \tag{4.32}$$

Define $s_t = \sum_{k=1}^{t} \|X_k\|_p$. Then we can obtain that

$$s_t \le s_{t+1} \le (\rho + \epsilon) s_t + o(t) \tag{4.33}$$

which implies $s(t) = o(t)$ by $\rho + \epsilon < 1$. Consequently,

$$\sum_{k=1}^{t} \|X_k\| = o(t) \tag{4.34}$$

can be obtained by the equivalence among norms. By using inequality $2xy \le \epsilon x^2 + \frac{1}{\epsilon} y^2$ and

$$\|X_{t+1}\|_p^2 \le \|A_t\|_p^2 \|X_t\|_p^2 + \|W_t\|_p^2 + 2\|A_t\|_p \|X_t\|_p \|W_t\|_p$$

we can similarly obtain $\sum_{k=1}^{t} \|X_k\|^2 = o(t)$. □

LEMMA 4.6. *Let system* (2.1) *satisfy Assumption* A1 *in Cases* I, II, III, *and* IV. *Then, for $i = 1, 2, \dots, N$, we have $\frac{\log r_i(t)}{t} = o(1)$ as $t \to \infty$ a.s.*

Proof. Putting (2.13) into (2.1), we have

$$\begin{aligned} x_i(t+1) &= -\hat{\gamma}_i(t)\bar{x}_i(t) + z_i(t) + \gamma_i \bar{x}_i(t) + w(t+1) \\ &= z_i(t) + \tilde{\gamma}_i(t)\bar{x}_i(t) + w(t+1). \end{aligned} \tag{4.35}$$

Denote

$$\begin{aligned} X(t) &= (x_1(t), x_2(t), \dots, x_N(t))^\tau, \\ Z(t) &= (z_1(t), z_2(t), \dots, z_N(t))^\tau, \\ \bar{X}(t) &= (\bar{x}_1(t), \bar{x}_2(t), \dots, \bar{x}_N(t))^\tau, \\ W(t+1) &= w(t+1)\mathbf{1} = (w(t+1), w(t+1), \dots, w(t+1))^\tau, \\ \tilde{\Gamma}(t) &= \text{diag}(\tilde{\gamma}_1(t), \tilde{\gamma}_2(t), \dots, \tilde{\gamma}_N(t))^\tau. \end{aligned} \tag{4.36}$$

Then we get

$$X(t+1) = Z(t) + \tilde{\Gamma}(t)\bar{X}(t) + W(t+1). \tag{4.37}$$

According to (2.2), we have

$$\bar{X}(t) = GX(t), \tag{4.38}$$

where the matrix $G = (g_{ij})$. Furthermore, we have

$$\bar{X}(t+1) = GX(t+1) = GZ(t) + G\tilde{\Gamma}(t)\bar{X}(t) + W(t+1). \tag{4.39}$$

By Corollary 4.1, we have $\tilde{\gamma}(t) \to 0$ as $t \to \infty$. Thus, $\tilde{\Gamma}(t) \to 0$.

By Assumption A1, we can deduce that

$$|w(t+1)| = O\left(t^{\delta}\right) \ a.s. \quad \forall \delta \in \left(\frac{1}{\beta}, \frac{1}{2}\right), \tag{4.40}$$

which can be obtained by the Borel–Cantelli–Levy lemma because

$$\sum_{t=1}^{\infty} P\left(w^2(t+1) \geq t^{2\delta} | \mathcal{F}_t\right) \leq \sum_{t=1}^{\infty} \frac{E\left[|w(t+1)|^{\beta} | \mathcal{F}_t\right]}{t^{\beta\delta}} < \infty \ a.s. \tag{4.41}$$

Define $S(t) = 1 + \sum_{k=1}^{t} \|\bar{X}(k)\|^2$. In the following we will consider four different cases, respectively:

(I) *Deterministic tracking.* In this case $Z(t) = z^*(t)\mathbf{1}$, where $z^*(t) = O(t^{\delta})$ is a sequence of deterministic signals. Obviously

$$\|GZ(t) + W(t+1)\| = O\left(t^{\delta}\right). \tag{4.42}$$

Then, by Lemma 4.4, we have

$$\left\|\bar{X}(t)\right\| = O\left(t^{\delta}\right), \ \frac{\log S(t)}{t} = O\left(\frac{(2\delta+1)\log t}{t}\right) = o(1). \tag{4.43}$$

(II) *Center-oriented tracking.* In this case $Z(t) = \frac{1}{N}E_N X(t)$, where E_N is a matrix of order N with all entries being 1. Then

$$\begin{aligned} X(t+1) &= \frac{1}{N}E_N X(t) + \tilde{\Gamma}(t)GX(t) + W(t+1) \\ &= \left(\frac{1}{N}E_N + \tilde{\Gamma}(t)G\right) X(t) + W(t+1). \end{aligned} \tag{4.44}$$

Obviously $\frac{1}{N}E_N + \tilde{\Gamma}(t)G \to \frac{1}{N}E_N$ as $t \to \infty$, noting that the spectral radius of $\frac{1}{N}E_N$ (a stochastic matrix) is 1, by Lemma 4.4, we have

$$\frac{\log S(t)}{t} = o(1). \tag{4.45}$$

(III) *Loose tracking.* In this case $Z(t) = \lambda\bar{X}(t)$, where $\lambda \in (0,1)$. Then

$$\bar{X}(t+1) = \left(\lambda G + G\tilde{\Gamma}(t)\right) \bar{X}(t) + W(t+1). \tag{4.46}$$

Since G is a stochastic nonnegative matrix, the spectral radius of G is 1. Noticing that $\lambda G + G\tilde{\Gamma}(t) \to \lambda G$ as $t \to \infty$, by Lemma 4.4, we have

$$\|\bar{X}(t)\| = O\left(t^{\delta}\right), \quad \frac{\log S(t)}{t} = O\left(\frac{(2\delta+1)\log t}{t}\right) = o(1). \tag{4.47}$$

(IV) *Tight tracking.* In this case $Z(t) = \bar{X}(t)$, where

$$\bar{X}(t+1) = \left(G + G\tilde{\Gamma}(t)\right)\bar{X}(t) + W(t+1). \tag{4.48}$$

Noticing that the spectral radius of G is 1 and $G + G\tilde{\Gamma}(t) \to G$ as $t \to \infty$, therefore, by Lemma 4.4, we have

$$\|X_t\| = o\left(t^{\delta}(1+\epsilon)^t\right) + O\left(t^{\delta}\right), \quad \frac{\log S(t)}{t} = o(1). \tag{4.49}$$

In all of the cases above, $\frac{\log S(t)}{t} = o(1)$. By $|\bar{x}_i(t)| = O(\|\bar{X}(t)\|)$ and according to the definitions of $\bar{r}_i(t)$ and $S(t)$, obviously we have $\bar{r}_i(t) = O(S(t))$. Hence, for $i = 1, 2, \ldots, N$, we have $\frac{\log \bar{r}_i(t)}{t} = o(1)$ as $t \to \infty$. □

LEMMA 4.7. *Suppose that Assumption* A1 *holds in Cases* I, II, III, *and* IV. *Then, in either case, for $i = 1, 2, \ldots, N$ and $m \geq 1$, $0 \leq d < m$, we have*

$$\begin{aligned} &\sum_{k=1}^{t} |\tilde{\gamma}_i(mk-d)\bar{x}_i(mk-d)|^2 = o(t) \ a.s., \\ &\sum_{k=1}^{t} |\tilde{\gamma}_i(mk-d)\bar{x}_i(mk-d)| = o(t) \ a.s. \end{aligned} \tag{4.50}$$

Proof. By (4.3) of Lemma 4.1,

$$\sum_{k=1}^{t} \bar{a}_i(k)\left[\tilde{\gamma}_i(k)\bar{x}_i(k)\right]^2 = O(\log \bar{r}_i(t)) \ a.s. \tag{4.51}$$

By Lemma 4.2, $\sum_{j=1}^{t} \bar{x}_i^2(j) \to \infty$ *a.s.* as $t \to \infty$. Then, by Lemma 4.3,

$$\bar{v}_i(k) = \bar{p}_i(k)\bar{x}_i^2(k) = \frac{\bar{x}_i^2(k)}{\sum_{j=1}^{k-1} \bar{x}_i^2(j)} = O(1) \ a.s. \tag{4.52}$$

Then we have

$$\begin{aligned} \sum_{k=1}^{t} \left[\tilde{\gamma}_i(k)\bar{x}_i(k)\right]^2 &= \sum_{k=1}^{t} \bar{a}_i(k)\left[\tilde{\gamma}_i(k)\bar{x}_i(k)\right]^2 \cdot \left[1 + \bar{v}_i(k)\right] \\ &= O\left(\log \bar{r}_i(t)\right) \ a.s. \end{aligned} \tag{4.53}$$

Together with Lemma 4.6, immediately we can get

$$\sum_{k=1}^{t} \left[\tilde{\gamma}_i(k)\bar{x}_i(k)\right]^2 = o(t) \ a.s. \tag{4.54}$$

Furthermore, by the Schwartz inequality,

$$\begin{aligned}\sum_{k=1}^{t}|\tilde{\gamma}_i(k)\bar{x}_i(k)| &\leq \left(\sum_{k=1}^{t}[\tilde{\gamma}_i(k)\bar{x}_i(k)]^2\right)^{\frac{1}{2}}\left(\sum_{k=1}^{t}1^2\right)^{\frac{1}{2}} \\ &= O\left(\sqrt{t\log\bar{r}_i(t)}\right) = o(t)\ a.s.\end{aligned} \tag{4.55}$$

Thus, the lemma is true for $m = 1$. As for $m > 1$, we need only replace t with mt. □

4.2. Proofs of theorems. Now we give the proofs of the theorems. Due to the couplings among agents, we adopt the basic methodology of local–global–local in our analysis.

Proof of Theorem 3.1. By (4.35), we have

$$x_i(t+1) - z_i(t) - w(t+1) = \tilde{\gamma}_i(t)\bar{x}_i(t). \tag{4.56}$$

Let $e_{ij}(t) \overset{\Delta}{=} x_i(t) - x_j(t)$, $\eta_i(t) = \tilde{\gamma}_i(t)\bar{x}_i(t)$. Then

$$e_{ij}(t+1) = [\eta_i(t) - \eta_j(t)] + [z_i(t) - z_j(t)]. \tag{4.57}$$

For convenience of later discussion, we introduce the following notations:

$$\begin{aligned} X(t) &= (x_1(t), x_2(t), \ldots, x_N(t))^\tau, \\ Z(t) &= (z_1(t), z_2(t), \ldots, z_N(t))^\tau, \\ \bar{X}(t) &= (\bar{x}_1(t), \bar{x}_2(t), \ldots, \bar{x}_N(t))^\tau, \\ G^\tau &= (\zeta_1, \zeta_2, \ldots, \zeta_N), \\ \mathbf{1} &= (1, 1, \ldots, 1)^\tau, \\ E(t) &= (e_{1N}(t), e_{2N}(t), \ldots, e_{N-1,N}(t), 0)^\tau, \\ \eta(t) &= (\eta_1(t), \eta_2(t), \ldots, \eta_N(t))^\tau. \end{aligned} \tag{4.58}$$

Case I. Here $z_i(t) = z^*(t)$, thus,

$$e_{ij}(t+1) = \eta_i(t) - \eta_j(t). \tag{4.59}$$

Consequently, by Lemma 4.7, we obtain that $(i \neq j)$

$$\sum_{k=1}^{t}|e_{ij}(k+1)|^2 = O\left(\sum_{k=1}^{t}\eta_i^2(t)\right) + O\left(\sum_{k=1}^{t}\eta_j^2(t)\right) = o(t), \tag{4.60}$$

and similarly $\sum_{k=1}^{t}|e_{ij}(k+1)| = o(t)$ also holds.

Case II. Here $z_i(t) = \bar{z}(t)$. The proof is similar to Case I.

Case III. Here $z_i(t) = \lambda\bar{x}_i(t) = \lambda\zeta_i^\tau X(t)$. Noting that $\zeta_i^\tau\mathbf{1} = 1$ for any i, we have

$$\zeta_i^\tau X(t) - \zeta_j^\tau X(t) = \zeta_i^\tau[X(t) - x_N(t)\mathbf{1}] - \zeta_j^\tau[X(t) - x_N(t)\mathbf{1}] = \zeta_i^\tau E(t) - \zeta_j^\tau E(t), \tag{4.61}$$

and, thus,

$$\begin{aligned} e_{ij}(t+1) &= [\eta_i(t) - \eta_j(t)] + \lambda[\bar{x}_i(t) - \bar{x}_j(t)] \\ &= [\eta_i(t) - \eta_j(t)] + \lambda[\zeta_i^\tau X(t) - \zeta_j^\tau X(t)] \\ &= [\eta_i(t) - \eta_j(t)] + \lambda[\zeta_i^\tau E(t) - \zeta_j^\tau E(t)]. \end{aligned} \tag{4.62}$$

Taking $j = N$ and $i = 1, 2, \ldots, N$, we can rewrite (4.62) into matrix form as

$$E(t+1) = [\eta(t) - \eta_N(t)\mathbf{1}] + \lambda[G - \mathbf{1}\zeta_N^\tau]E(t) = \lambda H E(t) + \xi(t), \tag{4.63}$$

where

$$H = G - G_N = G - \mathbf{1}\zeta_N^\tau, \;\; \xi(t) = \eta(t) - \eta_N(t). \tag{4.64}$$

By Lemma 4.7, we have

$$\sum_{k=1}^{t} \|\eta(k)\|^2 = o(t). \tag{4.65}$$

Therefore,

$$\sum_{k=1}^{t} \|\xi(k)\|^2 = o(t). \tag{4.66}$$

Now we prove that $\rho(H) \le 1$. In fact, for any vector v such that $v^\tau v = 1$, we have

$$\begin{aligned} |v^\tau H v| &= |v^\tau G v - v^\tau G_N v| \\ &\le \max\big(\lambda_{\max}(G)\|v\|^2 - \lambda_{\min}(G_N)\|v\|^2, \\ &\quad\ \lambda_{\max}(G_N)\|v\|^2 - \lambda_{\min}(G)\|v\|^2\big) \\ &\le \max\big(\|v\|^2, \lambda_{\max}(G_N)\|v\|^2\big) \\ &= 1 \end{aligned} \tag{4.67}$$

which implies that $\rho(H) \le 1$.

Finally, by (4.63), together with Lemma 4.5, we can immediately obtain

$$\sum_{k=1}^{t} \|E(k)\| = o(t), \quad \sum_{k=1}^{t} \|E(k)\|^2 = o(t). \tag{4.68}$$

Thus, for $i = 1, 2, \ldots, N-1$, as $t \to \infty$, we have proved

$$\frac{1}{t}\sum_{k=1}^{t} |e_{iN}(k)| \to 0, \quad \frac{1}{t}\sum_{k=1}^{t} [e_{iN}(k)]^2 \to 0. \qquad \square \tag{4.69}$$

Proof of Theorem 3.2. Case IV is similar to Case III. We need only prove that the spectral radius $\rho(H)$ of H is less than 1, i.e., $\rho(H) < 1$; then we can apply Lemma 4.5 like in Case III.

Consider the following linear system:

$$z(t+1) = Gz(t). \tag{4.70}$$

Noting that G is a stochastic matrix, then, by Assumption A2 and knowledge of the Markov chain, we have

$$\lim_{t\to\infty} G^t = \mathbf{1}\pi^\tau, \tag{4.71}$$

where π is the unique stationary probability distribution of the finite-state Markov chain with transmission probability matrix G. Therefore,

$$z(t) = G^t z_0 \to \mathbf{1}\pi^\tau z_0 = (\pi^\tau z_0)\mathbf{1} \tag{4.72}$$

which means that all elements of $z(t)$ converge to a same constant $\pi^\tau z_0$. Furthermore, let $z(t) = (z_1(t), z_2(t), \ldots, z_N(t))^\tau$ and $\nu(t) = (\nu_1(t), \nu_2(t), \ldots, \nu_{N-1}(t), 0)^\tau$, where $\nu_i(t) = z_i(t) - z_N(t)$ for $i = 1, 2, \ldots, N$. Then we can see that

$$\nu(t+1) = (G - G_N)\nu(t) = H\nu(t) \tag{4.73}$$

and $\lim_{t\to\infty} \nu(t) = 0$ for any initial values $\nu_i(0) \in \mathcal{R}$, $i = 1, 2, \ldots, N-1$. Obviously $\nu(t) = H^t\nu(0)$, and each entry in the Nth row of H^t is zero since each entry in the Nth row of H is zero. Thus, denote

$$H^t \triangleq \begin{pmatrix} H_0(t) & * \\ 0 & 0 \end{pmatrix}, \tag{4.74}$$

where $H_0(t)$ is an $(N-1) \times (N-1)$ matrix. Then, for $i = 1, 2, \ldots, N-1$, taking $\nu(0) = \mathbf{e}_i$, respectively, by $\lim\limits_{t\to\infty} \nu(t) = 0$ we easily know that the ith column of $H_0(t)$ tends to zero vector as $t \to \infty$. Consequently, we have

$$\lim_{t\to\infty} H_0(t) = 0, \tag{4.75}$$

and, consequently, each eigenvalue of $H_0(t)$ tends to zero too. By (4.74), eigenvalues of H^t are identical with those of $H_0(t)$ except for zero, and, thus, we obtain that

$$\lim_{t\to\infty} \rho\left(H^t\right) = 0 \tag{4.76}$$

which implies that

$$\rho(H) < 1. \tag{4.77}$$

This completes the proof of Theorem 3.2. □

5. Simulation examples.

In this section, we will illustrate several examples to verify the effectiveness of the decentralized LS-based adaptive controller presented in this paper.

Settings. The settings in all cases are listed in Table 5.1. In each figure, six subfigures are given, which illustrate the evolution process of states $x_i(t)$, control signals $u_i(t)$, noise sequence $w(t)$, estimates $\hat{\gamma}_i(t)$ of intensity γ_i of influence, mean $m_i^{(1)}(t)$ of absolute values of synchronization errors $\{e_i(t)\}$, and mean $m_i^{(2)}(t)$ of squared synchronization errors $\{e_i(t)\}$, respectively, where

$$\begin{aligned} e_i(t) &\triangleq x_i(t) - x_1(t), \\ m_i^{(1)}(t) &\triangleq \frac{1}{t}\sum_{k=1}^{t} |e_i(k)|, \\ m_i^{(2)}(t) &\triangleq \frac{1}{t}\sum_{k=1}^{t} |e_i(k)|^2. \end{aligned} \tag{5.1}$$

TABLE 5.1
Settings of simulations.

number N of agents	$N = 5$
time steps T	$T = 40$
noise sequence $\{w_t\}$	randomly taken from normal distribution $N(0;1)$
matrix G	randomly generated stochastic matrix
intensity γ_i of influence	randomly taken from interval $[0,1]$

Case I (deterministic tracking). $z_i(t) = z^*(t)$. Here we take $z^*(t) = 10\sin\frac{t}{3}$. A simulation in this case is shown in Figure 5.1.

Case II (center-oriented tracking). $z_i(t) = \bar{z}(t)$. A simulation in this case is shown in Figure 5.2.

Case III (loose tracking). $z_i(t) = \lambda\bar{x}_i(t)$. Here we take constant $\lambda = 0.7$. A simulation in this case is shown in Figure 5.3.

Case IV (tight tracking). $z_i(t) = \bar{x}_i(t)$. A simulation in this case is shown in Figure 5.4.

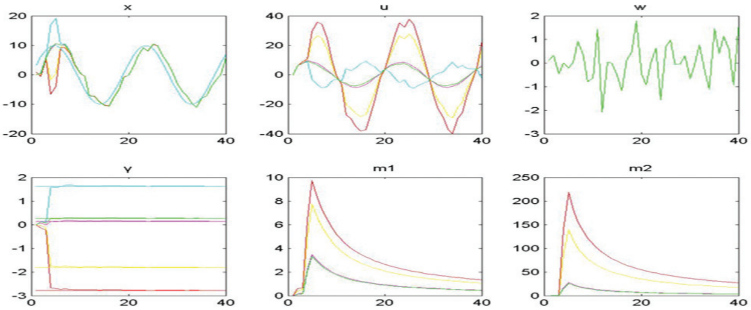


FIG. 5.1. *A simulation in Case* I *(deterministic tracking).*

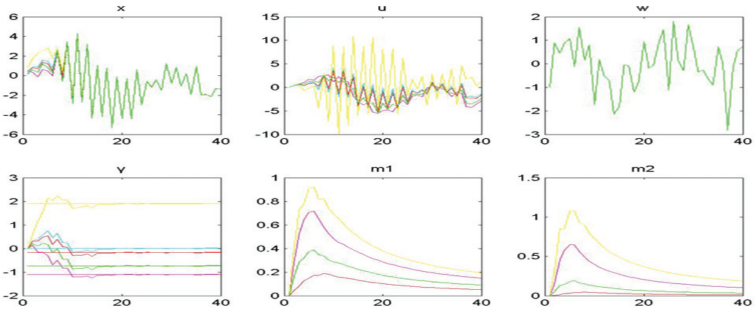


FIG. 5.2. *A simulation in Case* II *(center-oriented tracking).*

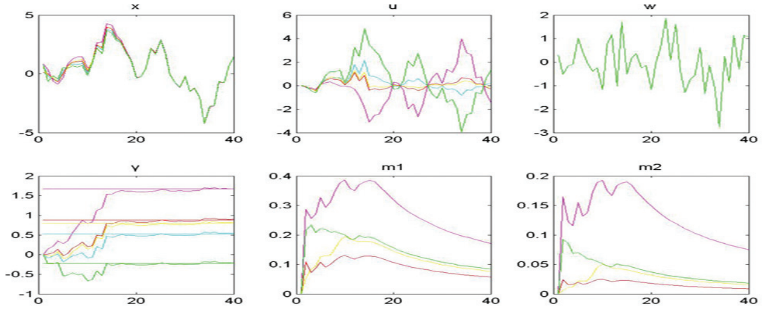


FIG. 5.3. *A simulation in Case* III *(loose tracking).*

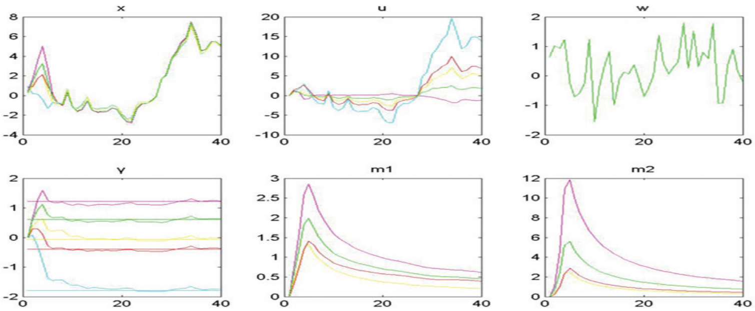


FIG. 5.4. *A simulation in Case* IV *(tight tracking).*

6. Summary. In this paper, for the sake of theoretical analysis, we first give a general framework on adaptive control problems for complex systems with uncertainties. The uncertainties may consist of noise disturbance, communication limits parametric coupling uncertainties among agents, and even internal parametric uncertainties or structural uncertainties in the agents themselves. Within this framework, we have studied the decentralized adaptive synchronization problem for a simple yet nontrivial discrete-time stochastic model, where agents can take effects on those agents in its local neighborhood, and we assume that the coupling effects are linear and unknown for each agent. For this simplest model with many agents, the following fundamental problem is considered: Can all agents achieve global synchronization while they are pursuing their local goals? Answers to this problem may help to understand deeply the relationship between local goals and the global goal in complex control systems.

Although the notion of *adaptive synchronization* has been investigated for continuous-time deterministic dynamical systems, we notice that, compared with continuous-time deterministic models, discrete-time stochastic models usually have different features and corresponding difficulties involved in theoretical analysis. To mathematically describe the synchronization phenomena in noisy systems, several novel definitions of synchronization in sense of mean are proposed for the study on complex systems with noise disturbance. By applying the new concepts of synchronization in sense of mean, we then formulate an adaptive synchronization problem mathematically for the considered discrete-time stochastic model. As to the local goals, we consider four different cases, including “deterministic tracking,” “center-oriented tracking,” “loose tracking,” and “tight tracking,” the first two of which correspond to cases with a hidden leader and the latter two of which correspond to leader-free cases.

Within our framework, since all agents are in the noisy environment and each agent does not know the coupling parameter (i.e., intensity γ_i of influence), each agent must use the proper learning algorithm and design its control law to reduce the effects of uncertainties in parameters and environment. In this contribution, agents are supposed to use local estimators and local controllers based on the LS algorithm to achieve their local goals since the LS algorithm is one of the mostly widely used recursive estimation algorithms in statistics, system identification, and adaptive control.

In the first three cases, we have proved that whatever the neighborhood relation (reflected in matrix G) is, global synchronization in sense of mean can be achieved by the decentralized LS-based learning and control algorithm. In the last case (“tight tracking”), we have proved that, under a weak condition on matrix G, global synchronization in sense of mean can also be achieved by the same algorithm. The condition imposed on matrix G cannot be weakened in general since it is necessary and sufficient even when there is no noise and no uncertainty in parameters γ_i. We should also remark that the assumption on noise sequence in these results is also very weak, since it allows unbounded noise including Gaussian white noise.

To the best knowledge of the author, the rigorous analysis for decentralized adaptive synchronization of the stochastic model in this paper is a first theoretical try in analyzing adaptive synchronization of a discrete-time stochastic complex dynamic network with uncertainties, which illustrates also our general framework on adaptive control of complex dynamic networks and several new concepts of synchronization for noisy systems. This contribution is still a starting point towards a comprehensive understanding for adaptive synchronization of discrete-time stochastic complex dynamic networks. Many related problems still remain to be solved in the future; for example, this paper considers only dynamical networks with fixed topology, while the study on general dynamical networks with time-varying topology may be more interesting and challenging.

Acknowledgment. The author is indebted to Prof. Lei Guo, who led me to the world of complex systems and gave me valuable advice on this paper.

REFERENCES

[1] V. Ahlers and A. Pikovsky, *Critical properties of the synchronization transition in space-time chaos*, Phys. Rev. Lett., 88 (2002), article 254101.

[2] K. Aström and B. Wittenmark, *Adaptive Control*, Addison-Wesley, Reading, MA, 1989.

[3] M. Barahona and L. M. Pecora, *Synchronization in small-world systems*, Phys. Rev. Lett., 89 (2002), article 054101.

[4] V. N. Belykh, I. V. Belykh, and M. Hasler, *Connection graph stability method for synchronized coupled chaotic systems.* I. *General approach*, Phys. D, 195 (2004), pp. 159–187.

[5] V. D. Blondel, J. M. Hendrickx, A. Olshevsky, and J. N. Tsitsiklis, *Convergence in the multiagent coordination, consensus, and flocking*, in Proceedings of the Joint 44th IEEE Conference on Decision and Control and European Control Conference, Seville, Spain, 2005.

[6] V. D. Blondel, J. M. Hendrickx, and J. N. Tsitsiklis, *On the* $2R$ *conjecture for multi-agent systems*, in Proceedings of the 2007 European Control Conference, Kos, Greece, 2007 European Union Control Association.

[7] J. Cao and J. Lu, *Adaptive synchronization of neural networks with or without time-varying delay*, Chaos, 16 (2006), article 013133.

[8] J. Cao, Z. Wang, and Y. Sun, *Synchronization in an array of linearly stochastically coupled networks with time delays*, Phys. A, 385 (2007), pp. 718–728.

[9] M. Cao, A. S. Morse, and B. D. O. Anderson, *Reaching a consensus in a dynamically changing environment: A graphical approach*, SIAM J. Control Optim., 47 (2008), pp. 575–600.

[10] H. F. Chen and L. Guo, *Identification and Stochastic Adaptive Control*, Birkhäuser, Boston, MA, 1991.

[11] T. Eren, B. D. O. Anderson, A. S. Morse, W. Whiteley, and P. B. Belhumeur, *Operations on rigid formations of autonomous agents*, Commun. Inf. Syst., 3 (2004), pp. 223–258.

[12] A. L. Fradkov and A. Y. Markov, *Adaptive synchronization of chaotic systems based on speed gradient method and passification*, IEEE Trans. Circuits Syst. I Fundam. Theory Appl., 44 (1997), pp. 905–912.

[13] A. L. Fradkov, I. V. Miroshnik, and V. O. Nikiforov, *Nonlinear and Adaptive Control of Complex Systems: Mathematics and Its Applications*, Kluwer Academic Publishers, Dordrecht, 2004.

[14] P. M. Gade and C.-K. Hu, *Synchronous chaos in coupled map lattices with small-world interactions*, Phys. Rev. E, 62 (2000), pp. 6409–6413.

[15] M. Gell-Mann, *The Quark and the Jaguar, Adventures in the Simple and the Complex*, W. H. Freeman, New York, 1994.

[16] G. Goodwin and K. Sin, *Adaptive Filtering, Prediction and Control*, Prentice–Hall, Englewood Cliffs, NJ, 1984.

[17] L. Guo, *Time-varing Stochastic Systems*, Ji Lin Science and Technology Press, Jilin, China, 1993 (in Chinese).

[18] L. Guo, *On critical stability of discrete-time adaptive nonlinear control*, IEEE Trans. Automat. Control, 42 (1997), pp. 1488–1499.

[19] L. Guo, *Exploring the maximum capability of adaptive feedback*, Internat. J. Adapt. Control Signal Process., 16 (2002), pp. 341–354.

[20] J. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, University of Michigan Press, Ann Arbor, MI, 1975.

[21] J. Holland, *Hidden Order: How Adaptation Builds Complexity*, Addison–Wesley, New York, 1996.

[22] X. Huang and J. Cao, *Generalized synchronization for delayed chaotic neural networks: A novel coupling scheme*, Nonlinearity, 19 (2006), pp. 2797–2811.

[23] P. A. Ioannou and J. Sun, *Robust Adaptive Control*, Prentice–Hall, Englewood, Cliffs, NJ, 1996.

[24] A. Jadbabaie, J. Lin, and A. S. Morse, *Coordination of groups of mobile autonomous agents using nearest neighbor rules*, IEEE Trans. Automat. Control, 48 (2003), pp. 998–1001.

[25] S. Jalan and R. E. Amritkar, *Self-organized and driven phase synchronization in coupled maps*, Phys. Rev. Lett., 90 (2003), article 014101.

[26] R. E. Kalman, *Design of self-optimizing control systems*, Trans. ASME J. Appl. Mech., 80 (1958), pp. 468–478.

[27] Y. D. Landau, *Adaptive Control: The Model Reference Approach*, Dekker, New York, 1979.

[28] X. Li and G. Chen, *Synchronization and desynchronization of complex dynamical networks: An engineering viewpoint*, IEEE Trans. Circuits Syst. I Regul. Pap., 50 (2003), pp. 1381–1390.

[29] Z. Li, G. Feng, and D. J. Hill, *Controlling complex dynamical networks with coupling delays to a desired orbit*, Phys. Lett. A, 359 (2006), pp. 42–46.

[30] J. Lin, A. S. Morse, and B. D. O. Anderson, *The multi-agent rendezvous problem*, in Proceedings of the 42nd IEEE Conference on Decision and Control, Maui, Hawaii, 2003, pp. 1508–1513.

[31] J. Lü and G. Chen, *A time-varying complex dynamical network model and its controlled synchronization criteria*, IEEE Trans. Automat. Control, 50 (2005), pp. 841–846.

[32] J. Lü, H. Leung, and G. Chen, *Complex dynamical networks: Modelling, synchronization and control*, Dyn. Contin. Discrete Impuls. Syst. Ser. B Appl. Algorithms, 11a (2004), pp. 70–77.

[33] H. B. Ma, *Finite-model adaptive control using LS-like algorithm*, Internat. J. Adapt. Control Signal Proces., 21 (2006), pp. 391–414.

[34] H. B. Ma, *Finite-model adaptive control using WLS-like algorithm*, Automatica J. IFAC, 43 (2007), pp. 677–684.

[35] H. B. Ma, *Further results on limitations to the capability of feedback*, Internat. J. Control, 81 (2008), pp. 21–42.

[36] H. B. Ma, *An "impossibility" theorem on a class of high-order discrete-time nonlinear control systems*, Systems Control Lett., 57 (2008), pp. 497–504.

[37] H. B. Ma, K. Y. Lum, and S. S. Ge, *Decentralized Åström-Wittenmark self-tuning regulator of a multi-agent uncertain coupled armax system*, in Proceedings of the 2007 IEEE Multi-conference on Systems and Control, Singapore, 2007.

[38] S. Mu, T. Chu, and L. Wang, *Coordinated collective motion in a motile particle group with a leader*, Phys. A, 351 (2005), pp. 211–226.

[39] R. Olfati-Saber, J. A. Fax, and R. M. Murray, *Consensus and cooperation in networked multi-agent systems*, Proc. IEEE, 95 (2007), pp. 215–233.

[40] H. Shi, L. Wang, and T. Chu, *Virtual leader approach to coordinated control of multiple mobile agents with asymmetric interactions*, Phys. D, 213 (2006), pp. 51–65.

[41] M. Time, F. Wolf, and T. Geisl, *Toplogical speed limits to network synchronization*, Phys. Rev. Lett., 92 (2004), article 074101.

[42] M. M. Waldrop, *Complexity: The Emerging Science at the Edge of Order and Chaos*, Simon & Schuster, New York, 1992.

[43] X. Wang and G. Chen, *Synchronization in scale-free dynamical networks: Robustness and fragility*, IEEE Trans. Circuits Syst. I Regul. Pap., 49 (2002), pp. 54–62.

[44] X. Wang and G. Chen, *Synchronization in small-world dynamical networks*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 12 (2002), p. 187.

[45] R. B. Williams, *Restricted complexity framework for nonlinear adaptive control in complex systems*, AIP Conf. Proc., 699 (2004), pp. 623–630.

[46] C. W. Wu and L. O. Chua, *Synchronization in an array of linearly coupled dynamical systems*, IEEE Trans. Circuits Syst. I, 42 (1995), pp. 430–447.

[47] L. L. Xie and L. Guo, *How much uncertainty can be dealt with by feedback?*, IEEE Trans. Automat. Control Regul. Pap., 45 (2000), pp. 2203–2217.

[48] F. Xue and L. Guo, *Stabilizability, uncertainty and the choice of sampling rate*, in Proceedings of the 39th IEEE Conference on Decision and Control, Sydney, 2000, p. 19.

[49] F. Xue, M. Y. Huang, and L. Guo, *Towards understanding the capability of adaptation for time-varying systems*, Automatica J. IFAC, 37 (2001), pp. 1551–1560.

[50] J. Yao, Z. H. Guan, and D. J. Hill, *Adaptive switching control and synchronization of chaotic systems with uncertainties*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 15 (2005), pp. 1–10.

[51] J. Yao, D. J. Hill, Z. H. Guan, and H. O. Wang, *Synchronization of complex dynamical networks with switching topology via adaptive control*, in Proceedings of 45th IEEE Conference on Decision and Control, San Diego, CA, 2006, pp. 2819–2824.

[52] M. Zhan, X. Wang, X. Gong, G. W. Wei, and C.-H. Lai, *Complete synchronization and generalized synchronization of one-way coupled time-delay systems*, Phys. Rev. E, 68 (2003), article 036208.

[53] Y. X. Zhang and L. Guo, *A limit to the capability of feedback*, IEEE Trans. Automat. Control, 47 (2002), pp. 687–692.

[54] J. Zhou, J. A. Lu, and J. Lü, *Adaptive synchronization of an uncertain complex dynamical network*, IEEE Trans. Automat. Control, 51 (2006), pp. 652–656.

SIAM J. Control Optim.
Vol. 48, No. 2, pp. 881–908

© 2009 Society for Industrial and Applied Mathematics

STRICT EFFICIENCY IN SET-VALUED OPTIMIZATION[*]

FABIÁN FLORES-BAZÁN[†] AND BIENVENIDO JIMÉNEZ[‡]

Abstract. In this paper, we develop the notion of a ϕ-minimizer (or strict efficiency) for a set-valued map. Its properties and relations with other similar notions are studied. In special circumstances, under suitable conditions, we prove that a point is a ϕ-minimizer of a vector function if and only if it is a ϕ-minimizer of a certain family of linear scalarizations. We also establish a characterization of strict efficiency through a nonlinear scalarization, which is a generalization of the Gerstewitz function defined on the power set of the image space. The final part is focused on minimizers of order one, and we provide several necessary or sufficient conditions (without convexity assumptions) through different kinds of derivatives as contingent, radial among others. Various illustrative examples showing the applicability of our results are also presented.

Key words. set-valued map, strict minimizer, strict efficient point, Gerstewitz function, optimality conditions, contingent derivative

AMS subject classifications. 90C29, 90C46, 49A52, 49B27, 90C31

DOI. 10.1137/07070139X

1. Introduction.

The notion of a strict minimizer (of order m) for a scalar function has proved to be very fruitful in optimization. Let us recall that given a normed space X, $f : X \to \mathbb{R}$ and $S \subset X$, a point $x_0 \in S$ is said to be a strict local minimizer of order m ($m \geq 1$ integer) for f over S if there exist a neighborhood U of x_0 and $\alpha > 0$ such that

$$f(x) \geq f(x_0) + \alpha \|x - x_0\|^m \quad \forall x \in S \cap U \setminus \{x_0\}. \tag{1.1}$$

Since the pioneering works by Auslender [4] and Studniarski [35] for scalar functions, the concept has also been extended and developed successfully in vector optimization (see, for example, [14, 22, 24, 28]), where it is already an established notion.

A variant, named ϕ-strict minimizer, was presented by Bednarczuk [6] for vector functions. Also the notion of minimizer of order one has been introduced for set-valued maps by Crespi, Ginchev, and Rocca [10] following the line of Ginchev, Guerraggio, and Rocca [14]. In this paper, we extend the notion of ϕ-strict minimizer to set-valued maps, following the line of Jiménez [22], in such a way that all of them are generalized in a unified manner.

The outline of the paper is as follows. In section 2, we introduce the notations. In section 3, the notion of ϕ-strict minimizer for a set-valued map is introduced, and

[*] Received by the editors August 28, 2007; accepted for publication (in revised form) October 13, 2008; published electronically February 25, 2009.
http://www.siam.org/journals/sicon/48-2/70139.html

[†] Departamento de Ingeniería Matemática, Facultad de Ciencias Físicas y Matemáticas, Universidad de Concepción, Casilla 160-C, Concepción, Chile (fflores@ing-mat.udec.cl). This research, for the first author, was supported in part by CONICYT-Chile through FONDECYT 107-0689, FONDAP and BASAL Projects CMM, Universidad de Chile and by Centro de Investigación en Ingeniería Matemática (CI2MA), Universidad de Concepción.

[‡] Departamento de Matemática Aplicada, E.T.S.I. Industriales, Universidad Nacional de Educación a Distancia, Calle Juan del Rosal, 12, 28040 Madrid, Spain (bjimenez@ind.uned.es). For the second author, this research was partially supported by the Ministerio de Educación y Ciencia (Spain), project MTM2006-02629. Part of this research was carried out during a stay of the second author in the Department of Mathematical Engineering, Universidad de Concepción (Chile), during November 2006 and was funded by project FONDECYT-Incentivo a la Cooperación Internacional 7060106. The author thanks the Department for its hospitality.

immediate properties are established. A structure theorem (Theorem 3.7) is proved for a vector-valued function; according to it, a point is a ϕ-strict minimizer of $f : X \to Y$ over a set if and only if the point is a ϕ-strict minimizer for a family of scalar functions and sets, each of these functions is the composition of f with a positive, continuous and linear functional, and the family of sets is a covering of the initial set. In this section, the composition with a continuous positively homogeneous functional is also analyzed.

In section 4, scalarization is tackled. For this aim, we use a generalization of the Gerstewitz function. This generalization is a function from the power set of Y (2^Y) into the real numbers, i.e., a set-to-point function. In this way we associate to the set-valued optimization problem an ordinary scalar optimization problem with the same feasible set, and we characterize different kinds of strict minimizers for the set-valued problem through different kinds of strict minimizers for the scalarized problem. Some consequences are obtained from the main result in this section.

In section 5, we pay attention on strict minimizers of order one, and several optimality conditions are established. A characterization is given for a global minimizer through the radial derivative. A sufficient condition and other necessary conditions are proved for a local minimizer through the Shi and contingent derivatives when the initial space is finite dimensional. In case such a space is of infinite dimension, we use the contingent derivative under the upper-semidifferentiability condition of the set-valued map to obtain a sufficient condition. Convexity of the set-valued map is not required. Comparisons with other results are made, and some illustrative examples are also provide.

2. Notations. Throughout the paper, X and Y are normed spaces. $B(x, \delta)$ denotes the open ball centered at $x \in X$ and radius $\delta > 0$, $d(x, S)$ is the distance from x to the set $S \subset X$. We denote by S^c, $\operatorname{cl} S$, $\operatorname{int} S$, $\operatorname{bd} S$ and $\operatorname{co} S$, the algebraic complement (i.e., $S^c = X \setminus S$), closure, interior, boundary, and convex hull of S, respectively. The cone generated by S is $\operatorname{cone} S = \{\alpha x :\ \alpha \geq 0,\ x \in S\}$.

The tangent cone to S at $x_0 \in \operatorname{cl} S$ is

$T(S, x_0) = \{v \in X :\ \exists t_n \to 0^+,\ \exists v_n \to v \text{ such that } x_0 + t_n v_n \in S\ \forall n \in \mathbb{N}\}$.

The expression $t_n \to 0^+$ means $t_n > 0\ \forall n$ and $t_n \to 0$.

We consider a convex cone $D \subset Y$, which defines a partial order on Y in the usual form: $y \leq_D z \Leftrightarrow z - y \in D\ \forall y, z \in Y$. We suppose that $0 \in D$ and D is a proper cone, that is, $\{0\} \neq D \neq Y$. We do not assume that D is pointed ($D \cap (-D) = \{0\}$) or closed.

We denote by 2^Y the power set of Y, i.e., the set of all subsets of Y. Given a set-valued map $F : X \to 2^Y$, we denote the graph and domain of F, respectively, by

$$\operatorname{Gr} F = \{(x, y) \in X \times Y :\ y \in F(x)\} \quad \text{and} \quad \operatorname{Dom} F = \{x \in X :\ F(x) \neq \emptyset\}.$$

If S is a subset of X, then $F(S) = \bigcup_{x \in S} F(x)$.

Definition 2.1. (a) *A point $y_0 \in A \subset Y$ is called an efficient (resp. a weak efficient) point of A and will be denoted by $y_0 \in \operatorname{Min}_D A$ (resp. $y_0 \in \operatorname{WMin}_D A$) if*

$$(A - y_0) \cap (-D) \subset D \quad (\textit{resp. } (A - y_0) \cap (-\operatorname{int} D) = \emptyset)$$

or equivalently, $(A - y_0) \cap (-D \setminus l(D)) = \emptyset$, where $l(D) = D \cap (-D)$. Of course, when the weak efficiency notion is considered, we suppose that $\operatorname{int} D \neq \emptyset$.

(b) *A point $y_0 \in A$ is called a strict efficient point of A, denoted by $y_0 \in \operatorname{Str}_D A$, if $(A - y_0) \cap (-D \setminus \{0\}) = \emptyset$.*

Note that

$$\operatorname{Str}_D A \subset \operatorname{Min}_D A \subset \operatorname{WMin}_D A,$$

and if D is pointed, then $\operatorname{Str}_D A = \operatorname{Min}_D A$. Let us observe that, in the literature for some authors, the set of efficient points is the set of strict efficient points even if D is not pointed. In our notion, in the set of efficient points, we admit indifferent points. For a different notion of strict minimality, we refer to [5].

A (nonnecessarily convex) subset B of D is a base of the cone D if $0 \notin B$ and for each $y \in D$, $y \neq 0$, there is a unique representation $y = \alpha b$, with $\alpha > 0$ and $b \in B$.

The topological dual space of Y is denoted by Y^*. The (positive) polar cone to D is $D^+ = \{\lambda \in Y^* : \ \langle \lambda, y\rangle \geq 0 \ \forall\, y \in D\}$.

Given a set-valued map $F : X \to 2^Y$ and a subset S of X, the following general vector optimization problem is considered:

$$\operatorname{Min}\{F(x) : \ x \in S\}. \tag{2.1}$$

Definition 2.2. *A point $(x_0, y_0) \in \operatorname{Gr} F$, with $x_0 \in S$, is said to be a local (resp. a local weak) minimizer of F over S, written $(x_0, y_0) \in \operatorname{LMin}_D(F, S)$ (resp. $(x_0, y_0) \in \operatorname{LWMin}_D(F, S)$), if there exists a neighborhood U of x_0 in X such that*

$$y_0 \in \operatorname{Min}_D F(U \cap S) \quad (\textit{resp. } y_0 \in \operatorname{WMin}_D F(U \cap S)),$$

i.e., $\forall x \in S \cap U$, $(F(x) - y_0) \cap (-D) \subset D$ (resp. $\forall x \in S \cap U$, $(F(x) - y_0) \cap (-\operatorname{int} D) = \emptyset$). The points of $\operatorname{LMin}_D(F, S)$ and $\operatorname{LWMin}_D(F, S)$ are also called local efficient solutions and local weak efficient solutions, respectively.

We will say that (x_0, y_0) is a global minimizer or global weak minimizer when we can choose $U = X$. The set of all global minimizers (resp. weak minimizers) is denoted by $\operatorname{Min}_D(F, S)$ (resp. $\operatorname{WMin}_D(F, S)$).

3. The notion of strict efficiency in optimization with set-valued maps.

In this section, we introduce the notion of ϕ-strict efficiency for set-valued maps that generalizes those due to Crespi, Ginchev, and Rocca [10], Bednarczuk [6], and Jiménez [22]. First, we recall the notion of strict minimizer.

Consider a set-valued map $F : X \to 2^Y$, a nonempty set $S \subset X$, and a proper convex cone $D \subset Y$.

Definition 3.1. *Let $x_0 \in S$. We say that a pair $(x_0, y_0) \in \operatorname{Gr} F$ is a strict local minimizer for F over S (or strict local efficient solution for problem (2.1)), denoted by $(x_0, y_0) \in \operatorname{Strl}_D(F, S)$, if* (i) *$y_0 \in \operatorname{Str}_D F(x_0)$ and* (ii) *there exists a neighborhood U of x_0 such that*

$$(F(x) - y_0) \cap (-D) = \emptyset \quad \forall x \in S \cap U \setminus \{x_0\}. \tag{3.1}$$

Notice that condition (i) $y_0 \in \operatorname{Str}_D F(x_0)$ means $(F(x_0) - y_0) \cap (-D \setminus \{0\}) = \emptyset$, which is a natural requirement for the pair (x_0, y_0) in line with (3.1).

Recall that $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ is said to be an admissible function if ϕ is nondecreasing, $\phi(0) = 0$, and $\phi(t) > 0$ for $t > 0$. Such a family of functions is denoted by $\mathcal{A}$.

Definition 3.2. *Let $\phi \in \mathcal{A}$ and $x_0 \in S$. We say that the pair $(x_0, y_0) \in \operatorname{Gr} F$ is a strict local minimizer with respect to ϕ for (2.1) (in short, a ϕ-strict local minimizer for F over S, or ϕ-strict local efficient solution), denoted by $(x_0, y_0) \in \operatorname{Strl}_D(F, S; \phi)$, if* (i) *$y_0 \in \operatorname{Str}_D F(x_0)$ and* (ii) *there exist a constant $\alpha > 0$ and a neighborhood U of x_0 such that*

$$(F(x) + D) \cap B(y_0, \alpha\phi(\|x - x_0\|)) = \emptyset \quad \forall x \in S \cap U \setminus \{x_0\}. \tag{3.2}$$

In particular, we will say that (x_0, y_0) is a strict local minimizer of order m (with $m > 0$), denoted by $(x_0, y_0) \in \mathrm{Strl}_D(F, S; m)$, if (x_0, y_0) is a strict local minimizer with respect to $\phi(t) = t^m$. This generalizes to set-valued maps the notion of strict minimum of order m given in Jiménez [22] and in Ginchev, Guerraggio, and Rocca [14] for single-valued functions $f : X \to Y$.

Condition (3.2) can be expressed in the following equivalent forms:

$$F(x) \subset (B(y_0, \alpha\phi(\|x - x_0\|)) - D)^c \quad \forall x \in S \cap U \setminus \{x_0\},$$

$$d(F(x) - y_0, -D) \geq \alpha\phi(\|x - x_0\|) \quad \forall x \in S \cap U \setminus \{x_0\}, \tag{3.3}$$

where the distance between two sets $A, B \in 2^Y$ is defined by

$$d(A, B) = \inf\{d(y, B) : \ y \in A\} = \inf\{\|y - z\| : \ y \in A, \ z \in B\}.$$

If we can choose $U = X$ as neighborhood of x_0 in Definitions 3.2 and 3.1, we will call (x_0, y_0) a ϕ-strict minimizer and strict minimizer, respectively, and they will be denoted by $\mathrm{Str}_D(F, S; \phi)$ and $\mathrm{Str}_D(F, S)$.

When F is a single-valued map, i.e., $F(x) = \{f(x)\}$, with $f : X \to Y$ a function, then we write $x_0 \in \mathrm{Strl}_D(f, S; \phi)$ instead of $(x_0, f(x_0)) \in \mathrm{Strl}_D(f, S; \phi)$ and similarly for the other sets of strict minimizers. Note that in this case, condition (i) in Definitions 3.1 and 3.2 is superfluous.

In particular, if $X = Y$ and $f(y) = y$, i.e., f is the identity ($f = \mathrm{id}$), we say that $y_0 \in S$ is a ϕ-strict efficient (resp. local efficient) point of S, denoted by $y_0 \in \mathrm{Str}_D(S; \phi)$ (resp. $y_0 \in \mathrm{Strl}_D(S; \phi)$), if $y_0 \in \mathrm{Str}_D(\mathrm{id}, S; \phi)$ (resp. $y_0 \in \mathrm{Strl}_D(\mathrm{id}, S; \phi)$). Similarly, it is defined that y_0 is a strict efficient (resp. local efficient) point of S, denoted by $y_0 \in \mathrm{Str}_D\, S$ (resp. $y_0 \in \mathrm{Strl}_D\, S$), if $y_0 \in \mathrm{Str}_D(\mathrm{id}, S)$ (resp. $y_0 \in \mathrm{Strl}_D(\mathrm{id}, S)$). Let us observe that this is coherent with Definition 2.1(b).

Bednarczuk [6, Definitions 3.3 and 3.4] gives a similar notion to Definition 3.2 for a function $f : X \to Y$. We extend her notion to a set-valued map as follows.

A pair $(x_0, y_0) \in \mathrm{Gr}\, F$ is said to be a ϕ-strict local minimizer in the sense of Bednarczuk, denoted $(x_0, y_0) \in \text{B-}\mathrm{Strl}_D(F, S; \phi)$, if (i) $y_0 \in \mathrm{Str}_D\, F(x_0)$ and (ii) there exists a neighborhood U of x_0 such that

$$(F(x) + D) \cap B(y_0, \phi(\|x - x_0\|)) = \emptyset \quad \forall\, x \in S \cap U \setminus \{x_0\}.$$

It is obvious the equality

$$\bigcup_{\alpha > 0} \text{B-}\mathrm{Strl}_D(F, S; \alpha\phi) = \mathrm{Strl}_D(F, S; \phi)$$

and consequently, $\text{B-}\mathrm{Strl}_D(F, S; \phi) \subset \mathrm{Strl}_D(F, S; \phi)$. The difference between these notions is slight but meaningful (for example, Proposition 3.4 is not true for the Bednarczuk notion).

The notion of strict local minimizer of order one for a set-valued map has been given by Crespi, Ginchev, and Rocca [10, Definition 2] (in a slightly different form), under the name of isolated minimizer. We prefer the name strict minimizer because in vector and in set-valued optimization, a strict minimizer is not, in general, isolated as in scalar optimization. Moreover, it is a natural generalization of the scalar case where the adjective strict is very usual (see, for example, [35, 38]).

In [10], the authors extend the so-called oriented distance to sets as follows:

$$\Delta(A, B) = \inf\{d(y, B) - d(y, Y \setminus B) : \ y \in A\}, \quad \text{for } A, B \in 2^Y,$$

and define the following: the pair $(x_0, y_0) \in \operatorname{Gr} F$ is a local isolated minimizer if there exist a neighborhood U of x_0 and $\alpha > 0$ such that

$$\Delta(F(x) - y_0, -D) \geq \alpha \|x - x_0\| \quad \forall x \in S \cap U. \tag{3.4}$$

Moreover, they require that y_0 be a proper efficient point of $F(x_0)$ in the sense of Henig instead of condition (i) in Definition 3.2. Note that (3.4) coincides with (3.3) for $\phi(t) = t$, since $\Delta(F(x) - y_0, -D) = d(F(x) - y_0, -D)$ whenever (3.4) holds.

Let us observe that if f is a scalar function, i.e., $f : X \to \mathbb{R}$ and $D = \mathbb{R}_+$, then

(i) $x_0 \in \operatorname{Strl}(f, S)$ if and only if there exists a neighborhood U of x_0 such that

$$f(x) > f(x_0) \quad \forall x \in S \cap U \setminus \{x_0\};$$

(ii) $x_0 \in \operatorname{Strl}(f, S; \phi)$ if and only if there exist $\alpha > 0$ and a neighborhood U of x_0 such that

$$f(x) \geq f(x_0) + \alpha \phi(\|x - x_0\|) \quad \forall x \in S \cap U \setminus \{x_0\}.$$

In particular, if $\phi(t) = t^m$, we obtain (1.1), i.e., the usual notion of strict local minimizer of order m [35, 38]. Note that for scalar functions, we write Str and Strl instead of $\operatorname{Str}_{\mathbb{R}_+}$ and $\operatorname{Strl}_{\mathbb{R}_+}$, respectively.

Now we present some immediate properties. Let $\phi, \varphi \in \mathcal{A}$. The following relationships are easy consequences of the above concepts:

(i) $\operatorname{Strl}_D(F, S; \phi) \subset \operatorname{Strl}_D(F, S)$.

(ii) $\operatorname{Strl}_D(F, S; \varphi) \subset \operatorname{Strl}_D(F, S; \phi)$ whenever $\varphi \geq \phi$ on $[0, \delta)$ for some $\delta > 0$.

(iii) $\operatorname{Str}_D(F, S; \phi) \subset \operatorname{Strl}_D(F, S; \phi)$.

(iv) $\operatorname{Str}_D(F, S) \subset \operatorname{Strl}_D(F, S)$.

(v) $\operatorname{Str}_D(F, S) \subset \operatorname{Min}_D(F, S)$ and $\operatorname{Strl}_D(F, S) \subset \operatorname{LMin}_D(F, S)$.

Consequently, any strict (local) notion implies (local) minimality. Other properties are as follows:

(vi) $(x_0, y_0) \in \operatorname{Strl}_D(F, S; \phi)$ if and only if there exists a neighborhood U of x_0 such that $(x_0, y_0) \in \operatorname{Str}_D(F, S \cap U; \phi)$. Similarly for (x_0, y_0) in $\operatorname{Strl}_D(F, S)$.

(vii) If $K \subset Y$ is another convex cone and $D \subset K$, then $\operatorname{Str}_K(F, S; \phi) \subset \operatorname{Str}_D(F, S; \phi)$, $\operatorname{Str}_K(F, S) \subset \operatorname{Str}_D(F, S)$ and similarly for the local notions.

The so-called profile map $F + D$ defined by $(F + D)(x) = F(x) + D$ is used very often in set-valued optimization and plays a crucial role in this paper. An important property of it is the following.

Proposition 3.3. *Let $F : X \to 2^Y$, $x_0 \in S \subset X$, $(x_0, y_0) \in \operatorname{Gr} F$, $\phi \in \mathcal{A}$ and assume that D is pointed. Then*

(a) $(x_0, y_0) \in \operatorname{Str}_D(F, S; \phi) \iff (x_0, y_0) \in \operatorname{Str}_D(F + D, S; \phi)$.

(b) $(x_0, y_0) \in \operatorname{Str}_D(F, S) \iff (x_0, y_0) \in \operatorname{Str}_D(F + D, S)$.

The same properties (a) *and* (b) *are true for the local notions.*

Proof. (a) Condition (3.2), with $U = X$, holds if and only if (3.2) holds for $F + D$ instead of F, since D is a convex cone. The implication $y_0 \in \operatorname{Str}_D(F(x_0) + D) \Rightarrow y_0 \in \operatorname{Str}_D F(x_0)$ is always true, and the converse also holds due to the pointedness of D. The proof of part (b) is also straightforward. □

The following proposition characterizes the situation where a point is a ϕ-strict minimizer in the general case, and Proposition 3.6 provides a necessary and sufficient condition in the Paretian case, i.e., $Y = \mathbb{R}^p$ and $D = \mathbb{R}^p_+$ is the nonnegative orthant.

We denote $\operatorname{Limsup}_{x \to x_0,\, x \in S} F(x)$ as the set of all cluster points of sequences $y_n \in F(x_n)$, with $x_n \to x_0$ and $x_n \in S \setminus \{x_0\}$.

Proposition 3.4. *Let $F : X \to 2^Y$, $x_0 \in S$, $y_0 \in \mathrm{Str}_D\, F(x_0)$, and $\phi \in \mathcal{A}$. Then, $(x_0, y_0) \in \mathrm{Strl}_D(F, S; \phi)$ if and only if*

$$0 \notin \operatorname*{Limsup}_{x\to x_0,\, x\in S} \frac{F(x) - y_0 + D}{\phi(\|x - x_0\|)}. \tag{3.5}$$

Proof. ($\Rightarrow$) Suppose that (3.5) is false. Then there exist sequences $x_n \in S \setminus \{x_0\}$, $y_n \in F(x_n)$, $d_n \in D$ such that $x_n \to x_0$ and $\lim_{n\to\infty} \frac{y_n - y_0 + d_n}{\phi(\|x_n - x_0\|)} = 0$. Then $\forall \varepsilon > 0$, $\exists n_0 = n_0(\varepsilon)$ such that $\forall n \geq n_0$ we have $x_n \in S$, $\|x_n - x_0\| < \varepsilon$, and $\|y_n - y_0 + d_n\| < \varepsilon\phi(\|x_n - x_0\|)$, that is, $y_n + d_n \in B(y_0, \varepsilon\phi(\|x_n - x_0\|))$.

By assumption, $(x_0, y_0) \in \mathrm{Strl}_D(F, S; \phi)$. Then, there exist $U = B(x_0, \delta)$ and $\alpha > 0$ such that (3.2) holds. Now, for $\varepsilon = \mathrm{Min}\{\delta, \alpha\}$, there exists $n_0 = n_0(\varepsilon)$ such that for each $n \geq n_0$, we have $x_n \in S \cap B(x_0, \delta)$ and

$$y_n + d_n \in B(y_0, \varepsilon\phi(\|x_n - x_0\|)) \subset B(y_0, \alpha\phi(\|x_n - x_0\|)),$$

which contradicts (3.2).

($\Leftarrow$) Suppose that $(x_0, y_0) \notin \mathrm{Strl}_D(F, S; \phi)$. Then $\forall \delta > 0$ and $\forall \alpha > 0$, $\exists x \in S \cap B(x_0, \delta) \setminus \{x_0\}$ $\exists y \in F(x)$ such that

$$(y + D) \cap B(y_0, \alpha\phi(\|x - x_0\|)) \neq \emptyset.$$

In particular, for $\delta = 1/n$ and $\alpha = 1/n$, there exist $x_n \in S \cap B(x_0, 1/n) \setminus \{x_0\}$, $y_n \in F(x_n)$, and $d_n \in D$ such that

$$y_n + d_n \in B\left(y_0, \tfrac{1}{n}\phi(\|x_n - x_0\|)\right),$$

that is,

$$\frac{\|y_n + d_n - y_0\|}{\phi(\|x_n - x_0\|)} < \frac{1}{n},$$

and so $\frac{y_n + d_n - y_0}{\phi(\|x_n - x_0\|)} \to 0$, which contradicts (3.5). □

Remark 3.5. A direct consequence of Proposition 3.4 is the following necessary condition for a point to be a ϕ-minimizer:

$$(x_0, y_0) \in \mathrm{Strl}_D(F, S; \phi) \;\Rightarrow\; \left(\operatorname*{Limsup}_{x\to x_0,\, x\in S} \frac{F(x) - y_0}{\phi(\|x - x_0\|)}\right) \cap (-D) = \emptyset.$$

The next proposition, the proof of which is similar to that of Proposition 3.5 in Jiménez [22] and so it is omitted, provides a useful characterization of local strict efficiency in the Paretian case. Set $\overline{\mathbb{R}}^p_+ = [0, +\infty]^p$.

Proposition 3.6. *Let $F : X \to 2^{\mathbb{R}^p}$, $x_0 \in S$, $y_0 \in \mathrm{Str}_{\mathbb{R}^p_+} F(x_0)$, and $\phi \in \mathcal{A}$. Then,*

$$(x_0, y_0) \in \mathrm{Strl}_{\mathbb{R}^p_+}(F, S; \phi) \iff \left(\operatorname*{Limsup}_{x\to x_0,\, x\in S} \frac{F(x) - y_0}{\phi(\|x - x_0\|)}\right) \cap \left(-\overline{\mathbb{R}}^p_+\right) = \emptyset.$$

The following theorem establishes a characterization of strict efficiency for a function under the assumption $\mathrm{int} D \neq \emptyset$ when efficiency is with respect to ϕ. The characterization reduces the strict efficiency of a vector-valued function f to the strict efficiency of a family of scalar functions obtained by means of the composition of f with a positive continuous linear functional (i.e., a functional of D^+).

Let D be a proper closed convex cone of Y with $\operatorname{int} D \neq \emptyset$. According to Lemma 2.2.17 in [16], the cone D^+ has a weak*-compact convex base Λ. Recall that a point $\lambda_0 \in \Lambda$ is an extremal point of a (convex) set Λ if there exist no different points $\lambda_1, \lambda_2 \in \Lambda$, and $t \in (0,1)$ such that $\lambda_0 = t\lambda_1 + (1-t)\lambda_2$.

THEOREM 3.7. *Let $f : X \to Y$ be a function, $D \subset Y$ be a proper closed convex cone, $x_0 \in S \subset X$, and $\phi \in \mathcal{A}$.*

(a) *Assume that* $\operatorname{int} D \neq \emptyset$*, let Λ be a weak*-compact convex base of D^+, and let Q be the set of extremal points of Λ. Then, $x_0 \in \operatorname{Str}_D(f, S; \phi)$ if and only if there exist $\rho > 0$ and a covering $\{V_\lambda : \lambda \in Q\}$ of $S \setminus \{x_0\}$ such that*

$$(3.6) \qquad \langle \lambda, f(x) \rangle > \langle \lambda, f(x_0) \rangle + \rho\phi(\|x - x_0\|) \quad \forall x \in V_\lambda \ \forall \lambda \in Q.$$

(b) *Let $Q \subset D^+ \setminus \{0\}$ and assume that $D^+ = \operatorname{cl} \operatorname{cone} \operatorname{co} Q$. Then $x_0 \in \operatorname{Str}_D(f, S)$ if and only if there exists a covering $\{V_\lambda : \lambda \in Q\}$ of $S \setminus \{x_0\}$ such that*

$$(3.7) \qquad \langle \lambda, f(x) \rangle > \langle \lambda, f(x_0) \rangle \quad \forall x \in V_\lambda \ \forall \lambda \in Q.$$

Proof. (a) ($\Rightarrow$) By assumption, there exists $\alpha > 0$ such that

$$(3.8) \qquad (f(x) - f(x_0) + D) \cap B(0, \alpha\phi(\|x - x_0\|)) = \emptyset \quad \forall x \in S \setminus \{x_0\}.$$

Let $e \in \operatorname{int} D$ be a fixed point, and let β be a fixed positive number that will be determined later. Let $\alpha_0 = \inf_{\lambda \in \Lambda} \langle \lambda, e \rangle$. The infimum is attained because Λ is weak*-compact and is attained at an extremal point of Λ, i.e., at a point of Q. Moreover, $\alpha_0 > 0$ because $\langle \lambda, e \rangle > 0 \ \forall \lambda \in \Lambda \subset D^+ \setminus \{0\}$, and so $\alpha_0 = \inf_{\lambda \in Q} \langle \lambda, e \rangle > 0$.

For each $\lambda \in Q$, we define

$$(3.9) \qquad V_\lambda = \{x \in X : \ \langle \lambda, f(x) \rangle > \langle \lambda, f(x_0) \rangle + \beta\phi(\|x - x_0\|)\alpha_0\}.$$

Let us see that

$$(3.10) \qquad S \setminus \{x_0\} \subset \bigcup_{\lambda \in Q} V_\lambda.$$

Pick any $x \in S \setminus \{x_0\}$ and assume that $x \notin V_\lambda \ \forall \lambda \in Q$. Then $\langle \lambda, f(x) \rangle \leq \langle \lambda, f(x_0) \rangle + \beta\phi(\|x - x_0\|)\alpha_0 \ \forall \lambda \in Q$, and as $\alpha_0 \leq \langle \lambda, e \rangle \ \forall \lambda \in Q$, we have

$$\langle \lambda, f(x_0) + \beta\phi(\|x - x_0\|)e - f(x) \rangle \geq 0 \quad \forall \lambda \in Q.$$

The same is true $\forall \lambda \in D^+$ and by the bipolar theorem,

$$d := f(x_0) + \beta\phi(\|x - x_0\|)e - f(x) \in D.$$

Hence,

$$(3.11) \qquad f(x) - f(x_0) + d = \beta\phi(\|x - x_0\|)e.$$

By choosing $\beta = \frac{\alpha}{2\|e\|}$, we get $\|\beta\phi(\|x - x_0\|)e\| < \alpha\phi(\|x - x_0\|)$. But then (3.11) contradicts (3.8). So, from (3.9) and (3.10), we deduce that (3.6) is satisfied with $\rho = \beta\alpha_0 = \alpha\alpha_0/(2\|e\|)$.

($\Leftarrow$) Let us prove that (3.8) holds for a suitable $\alpha > 0$. We define $\alpha_1 = \sup_{\lambda \in \Lambda} \langle \lambda, e \rangle$. This number is attained because Λ is a weak*-compact convex base of D^+, and, moreover, the supremum is attained at an extremal point. So it is clear that $0 < \alpha_1 = \max_{\lambda \in Q}\langle \lambda, e\rangle < +\infty$ and then $\alpha_1^{-1}\langle \lambda, e\rangle \leq 1\ \forall \lambda \in Q$. Hence, from (3.6) it follows that

$$\langle \lambda, f(x)\rangle > \langle \lambda, f(x_0)\rangle + \rho\phi(\|x - x_0\|) \geq \langle \lambda, f(x_0)\rangle + \alpha_1^{-1}\rho\phi(\|x - x_0\|)\langle \lambda, e\rangle \tag{3.12}$$

for each $x \in V_\lambda$ and for every $\lambda \in Q$. Now suppose that for all $\alpha > 0$, (3.8) is false. Then there exist $x' \in S \setminus \{x_0\}$ and $d \in D$ such that

$$f(x') - f(x_0) + d \in B(0, \alpha\phi(\|x' - x_0\|)). \tag{3.13}$$

As $0 \in \operatorname{int}(e - D)$ and D is a cone, there exists $k > 0$ such that $B(0, 1) \subset ke - D$ and consequently,

$$B(0, \alpha\phi(\|x' - x_0\|)) \subset k\alpha\phi(\|x' - x_0\|)e - D.$$

In view of (3.13), there is $d' \in D$ such that

$$f(x') - f(x_0) = k\alpha\phi(\|x' - x_0\|)e - (d + d').$$

As $x' \in S \setminus \{x_0\} \subset \cup_{\lambda \in Q} V_\lambda$, there exists $\bar{\lambda} \in Q$ such that $x' \in V_{\bar{\lambda}}$. As $Q \subset D^+$ and $d + d' \in D$, it follows that

$$\left\langle \bar{\lambda}, f(x')\right\rangle - \left\langle \bar{\lambda}, f(x_0)\right\rangle = k\alpha\phi(\|x' - x_0\|)\left\langle \bar{\lambda}, e\right\rangle - \left\langle \bar{\lambda}, d + d'\right\rangle \leq k\alpha\phi(\|x' - x_0\|)\left\langle \bar{\lambda}, e\right\rangle.$$

By choosing $\alpha = \alpha_1^{-1}\rho k^{-1}$, we obtain a contradiction to (3.12). Let us observe that α does not depend on λ.

(b) ($\Rightarrow$) For each $\lambda \in Q$, we define

$$V_\lambda = \{x \in X : \ \langle \lambda, f(x)\rangle > \langle \lambda, f(x_0)\rangle\}.$$

Let us see that (3.10) is satisfied. Pick any $x \in S \setminus \{x_0\}$ and assume that $x \notin V_\lambda\ \forall \lambda \in Q$. Then $\langle \lambda, f(x) - f(x_0)\rangle \leq 0\ \forall \lambda \in Q$, and the same inequality is true $\forall \lambda \in \operatorname{cl}\operatorname{cone}\operatorname{co} Q = D^+$. By the bipolar theorem, $f(x) - f(x_0) \in -D$, which contradicts the hypothesis. So (3.10) holds, and (3.7) is satisfied by the definition of V_λ.

($\Leftarrow$) Assume that $f(x) - f(x_0) \in -D$ for some $x \in S \setminus \{x_0\}$. Then $\langle \lambda, f(x)\rangle \leq \langle \lambda, f(x_0)\rangle\ \forall \lambda \in D^+$ (in particular, $\forall \lambda \in Q$). But this contradicts the assumption. □

Remark 3.8. (1) It is clear that Q is finite if and only if D is polyhedral.

(2) Many of the sets V_λ may be empty. As a matter of fact, only the sets V_λ satisfying $V_\lambda \cap S \neq \emptyset$ are of interest. If f and ϕ are continuous, then the sets V_λ can be chosen open.

(3) Note that expression (3.6) says that for each $\lambda \in Q$, x_0 is a strict minimizer of the scalar function $\lambda \circ f$ over $V_\lambda \cup \{x_0\}$ with respect to ϕ and with the same constant ρ. The converse is also true. But if ρ depends on λ, i.e., $\rho = \rho_\lambda$ on each V_λ, we can only ensure the converse if $\inf_{\lambda \in Q} \rho_\lambda > 0$. In particular, this is true if Q is finite.

(4) Part (b) can be rewritten as follows: $x_0 \in \operatorname{Str}_D(f, S)$ if and only if there exists a covering $\{V_\lambda : \ \lambda \in Q\}$ of $S \setminus \{x_0\}$ such that $x_0 \in \operatorname{Str}(\lambda \circ f, V_\lambda \cup \{x_0\})\ \forall \lambda \in Q$. Notice that as a set Q satisfying the requirement in part (b) is $\{\lambda \in D^+ : \ \|\lambda\| = 1\}$ or the set of extremal points of a convex base for D^+. A convex base exists if $\operatorname{qint} D := \{y \in Y : \ \langle \lambda, y\rangle > 0\ \forall \lambda \in D^+ \setminus \{0\}\} \neq \emptyset$; in such a situation a convex base is the set $\{\lambda \in D^+ : \ \langle \lambda, y\rangle = 1\}$ for $y \in \operatorname{qint} D$.

(5) If Q is finite, i.e., $Q = \{\lambda_i \in D^+ : \ i = 1, \dots, p\}$, then Theorem 3.7 is also true for the local notions. Consequently, part (a) can be written as follows:

$$\begin{aligned} x_0 \in \mathrm{Str}_D(f, S; \phi) \iff & \exists V_i \subset X, \ i = 1, \dots, p \text{ such that } S \setminus \{x_0\} \subset \cup_{i=1}^p V_i \text{ and} \\ & x_0 \in \mathrm{Str}(\lambda_i \circ f, V_i \cup \{x_0\}; \phi), \\ x_0 \in \mathrm{Strl}_D(f, S; \phi) \iff & \exists V_i \subset X, \ i = 1, \dots, p \text{ such that } S \setminus \{x_0\} \subset \cup_{i=1}^p V_i \text{ and} \\ & x_0 \in \mathrm{Strl}(\lambda_i \circ f, V_i \cup \{x_0\}; \phi). \end{aligned}$$

(6) In particular, if we choose $Y = \mathbb{R}^p$, $D = \mathbb{R}^p_+$, $\phi(t) = t^m$, and $Q = \{\lambda_1, \dots, \lambda_p\}$ as the canonical basis of $\mathbb{R}^p$, then we obtain Theorem 3.7 in [22].

Theorem 3.7 can be viewed as a sort of generalization to the case of strict minimizers of a family of scalarizations considered, for instance, in Paragraph 3.4.3 of [33].

To illustrate the above results, we give an example.

Example 3.9. (a) Let $f : \mathbb{R}^2 \to \mathbb{R}^3$ be given by $f(x) = (x_2, x_1^2, x_2^3)$ for $x = (x_1, x_2) \in \mathbb{R}^2$, $S = \mathbb{R}^2$, $x_0 = (0, 0)$, and $D = \{(y_1, y_2, y_3) \in \mathbb{R}^3 : \ y_1 - y_2 \geq 0, \ y_2 \geq 0, \ -y_2 - y_3 \geq 0\}$. It is clear that the set of extremal points of a convex base of D^+ is $Q = \{\lambda_1 = (1, -1, 0), \ \lambda_2 = (0, 1, 0), \ \lambda_3 = (0, -1, -1)\}$. Choosing as covering of $S \setminus \{x_0\}$ the sets $V_1 = \{(0, x_2) : \ x_2 > 0\}$, $V_2 = \{(x_1, x_2) : \ x_1 \neq 0\}$, and $V_3 = \{(0, x_2) : \ x_2 < 0\}$, one has that $x_0 \in \mathrm{Str}(\lambda_i \circ f, V_i \cup \{x_0\})$ for $i = 1, 2, 3$. Therefore, by Theorem 3.7 (taking into account Remark 3.8(4)), $x_0 \in \mathrm{Str}_D(f, S)$.

(b) With the same data, but now $f(x) = (x_2, x_1^2, x_2 + 2x_1^2)$. Choosing as covering of $S \setminus \{x_0\}$ the sets $V_1 = \{x : \ x_2 > |x_1|\}$, $V_2 = \{x : \ |x_2| < 2|x_1|\}$, and $V_3 = \{x : \ x_2 < -|x_1|\}$, one has: $x_0 \in \mathrm{Strl}(\lambda_1 \circ f, V_1 \cup \{x_0\}; 1)$, $x_0 \in \mathrm{Strl}(\lambda_2 \circ f, V_2 \cup \{x_0\}; 2)$, and $x_0 \in \mathrm{Strl}(\lambda_3 \circ f, V_3 \cup \{x_0\}; 1)$ (in order to check these claims, Theorems 6.3 and 6.4 in [17, Chapter 4] can be applied). Hence, $x_0 \in \mathrm{Strl}(\lambda_i \circ f, V_i \cup \{x_0\}; 2)$ for $i = 1, 2, 3$, and so $x_0 \in \mathrm{Strl}_D(f, S; 2)$ by Theorem 3.7 taking into account Remark 3.8(5).

Let us observe that in part (a), we have that $x_0 \notin \mathrm{Strl}_D(f, S; 2)$. Indeed, if we choose $t_n \to 0^+$, $x_n = (t_n^2, -t_n)$, $d_n = (t_n, 0, 0) \in D$, then $\lim_{n\to\infty} \frac{f(x_n) - f(x_0) + d_n}{\|x_n - x_0\|^2} = 0$, and the conclusion follows from Proposition 3.4 with $\phi(t) = t^2$.

Next, we study the composition. Let $\bar{Y}$ be a normed space, and let $\bar{D} \subset \bar{Y}$ be a proper convex cone that provides to $\bar{Y}$ a partial order. A function $\psi : Y \to \bar{Y}$ is said to be increasing if $\forall y, y' \in Y$, $y' \in y + D$ implies $\psi(y') \in \psi(y) + \bar{D}$.

Proposition 3.10. *Let $F : X \to 2^Y$, $x_0 \in S \subset X$, $(x_0, y_0) \in \mathrm{Gr}\, F$, and let $\psi : Y \to \bar{Y}$ be an increasing and positively homogeneous function which is continuous at* 0 *and either $\bar{D}$-convex or $\bar{D}$-concave. If $(x_0, \psi(y_0)) \in \mathrm{Str}_{\bar{D}}(\psi \circ F, S; \phi)$ and*

$$\psi^{-1}(\psi(y_0)) \cap F(x_0) = \{y_0\}, \tag{3.14}$$

then $(x_0, y_0) \in \mathrm{Str}_D(F, S; \phi)$.

Proof. By assumption, there exists $\alpha > 0$ such that

$$\big(\psi(F(x)) + \bar{D} - \psi(y_0)\big) \cap B(0, \alpha\phi(\|x - x_0\|)) = \emptyset \quad \forall x \in S \setminus \{x_0\}. \tag{3.15}$$

Since ψ is continuous at 0, given $\varepsilon = 1$ there is $\delta > 0$ such that $\psi(B(0, \delta)) \subset B(0, 1)$. As ψ is positively homogeneous, it follows that $\psi(B(0, 1)) \subset B(0, k)$, with $k = \delta^{-1}$, and therefore,

$$\psi(B(0, r)) \subset B(0, kr) \quad \forall r > 0. \tag{3.16}$$

Let $\beta = \alpha k^{-1}$, and let us prove that

$$(F(x) + D - y_0) \cap B(0, \beta\phi(\|x - x_0\|)) = \emptyset \quad \forall x \in S \setminus \{x_0\}.$$

Suppose that this is false, i.e., there exist $x \in S \setminus \{x_0\}$, $y \in F(x)$, and $d \in D$ such that

$$y' := y + d - y_0 \in B(0, \beta\phi(\|x - x_0\|)). \tag{3.17}$$

Then, from (3.16) we deduce that $\psi(y') \in B(0, k\beta\phi(\|x - x_0\|))$.

Case (A). ψ is $\bar{D}$-convex. From (3.17), $y = y_0 + y' - d$. As ψ is increasing and $\bar{D}$-convex, it follows that

$$\psi(y_0 + y' - d) \leq \psi(y_0 + y') \leq \psi(y_0) + \psi(y'),$$

and so $\psi(y) \leq \psi(y_0) + \psi(y')$. This implies that there is $\bar{d} \in \bar{D}$ such that $\psi(y) + \bar{d} = \psi(y_0) + \psi(y')$, i.e.,

$$\psi(y) - \psi(y_0) + \bar{d} = \psi(y') \in B(0, k\beta\phi(\|x - x_0\|)) = B(0, \alpha\phi(\|x - x_0\|)),$$

which contradicts (3.15).

Case (B). ψ is $\bar{D}$-concave. From (3.17), $y_0 = y - y' + d$. As ψ is increasing and $\bar{D}$-concave, it follows that

$$\psi(y - y' + d) \geq \psi(y - y') \geq \psi(y) + \psi(-y'),$$

and so $\psi(y_0) \geq \psi(y) + \psi(-y')$. This implies that there is $\bar{d} \in \bar{D}$ such that $\psi(y_0) = \psi(y) + \psi(-y') + \bar{d}$, i.e.,

$$\psi(y) - \psi(y_0) + \bar{d} = -\psi(-y') \in B(0, k\beta\phi(\|x - x_0\|)) = B(0, \alpha\phi(\|x - x_0\|)),$$

which contradicts (3.15).

Finally, let us see that $y_0 \in \operatorname{Str}_D F(x_0)$. Indeed, assume that $y_0 \notin \operatorname{Str}_D F(x_0)$, then there exists $y \in F(x_0)$ such that $y - y_0 = -d \neq 0$, with $d \in D$. As ψ is increasing, $\psi(y) = \psi(y_0) - \bar{d}$ for some $\bar{d} \in \bar{D}$. If $\bar{d} = 0$, by (3.14) it follows that $y = y_0$, which is a contradiction. If $\bar{d} \neq 0$, the condition $\psi(y) - \psi(y_0) = \bar{d}$ contradicts the condition $\psi(y_0) \in \operatorname{Str}_{\bar{D}} \psi(F(x_0))$, which is true by assumption. □

Remark 3.11. (1) Condition (3.14) can be removed if ψ is strictly increasing, i.e., $\forall y, y' \in Y$, $y' \in y + D \setminus \{0\}$ implies $\psi(y') \in \psi(y) + \bar{D} \setminus \{0\}$. Observe that condition (3.14) is only used to prove that $y_0 \in \operatorname{Str}_D F(x_0)$, and it is always satisfied if F is single-valued.

(2) Some examples of functions ψ satisfying the assumptions in Proposition 3.10 are the following: (a) for $\bar{Y} = \mathbb{R}$ and $D = \mathbb{R}_+$, $\psi = \lambda \in D^+$, $\psi(y) = \max_{i=1,\dots,l}\langle \lambda_i, y\rangle$ or $\psi(y) = \min_{i=1,\dots,l}\langle \lambda_i, y\rangle$, with $\lambda_i \in D^+$; (b) for a general $\bar{Y}$ and $\bar{D}$, $\psi : Y \to \bar{Y}$ a positive ($\psi(D) \subset \bar{D}$) continuous linear function, in particular, $\psi(y) = \langle \lambda, y\rangle \bar{d}$, with $\bar{d} \in \bar{D}$ and $\lambda \in D^+$.

In the next result, we prove the converse of Proposition 3.10 under more restrictive conditions.

Proposition 3.12. *Let Y, $\bar{Y}$ be Banach spaces, $\bar{D} \subset \bar{Y}$ a proper convex cone, $F : X \to 2^Y$, $x_0 \in S \subset X$, $(x_0, y_0) \in \operatorname{Gr} F$, and $\psi : Y \to \bar{Y}$ an onto continuous linear function. Assume that*

$$D = \left\{y \in Y :\ \psi(y) \in \bar{D}\right\}. \tag{3.18}$$

Then $(x_0, y_0) \in \operatorname{Str}_D(F, S; \phi) \ \Rightarrow\ (x_0, \psi(y_0)) \in \operatorname{Str}_{\bar{D}}(\psi \circ F, S; \phi)$.

Proof. By assumption, there exists $\alpha > 0$ such that

$$(F(x) + D - y_0) \cap B(0, \alpha\phi(\|x - x_0\|)) = \emptyset \quad \forall x \in S \setminus \{x_0\}. \tag{3.19}$$

As ψ is an onto continuous linear function, by the open mapping theorem, there exists $\delta > 0$ such that $B(0,\delta) \subset \psi(B(0,1))$. By linearity

$$B(0,r) \subset \psi(B(0,kr)) \quad \forall r > 0, \tag{3.20}$$

with $k = \delta^{-1}$. Set $\beta = \alpha k^{-1}$, and let us prove that

$$\big((\psi \circ F)(x) + \bar{D} - \psi(y_0)\big) \cap B(0, \beta\phi(\|x - x_0\|)) = \emptyset \quad \forall x \in S \setminus \{x_0\}.$$

Suppose that this is false, then there exist $x \in S \setminus \{x_0\}$, $y \in F(x)$, and $\bar{d} \in \bar{D}$ such that

$$\psi(y) + \bar{d} - \psi(y_0) \in B(0, \beta\phi(\|x - x_0\|)). \tag{3.21}$$

As ψ is onto, $\bar{d} = \psi(d)$ for some $d \in Y$ and by assumption (3.18), $d \in D$. Replacing in (3.21) and using (3.20), we have that

$$\psi(y + d - y_0) \in B(0, \beta\phi(\|x - x_0\|)) \subset \psi(B(0, k\beta\phi(\|x - x_0\|))).$$

From here, $y + d - y_0 = b + u$ for some $b \in B(0, k\beta\phi(\|x - x_0\|))$ and $u \in \ker\psi$, and therefore, $y + d - u - y_0 = b$, which contradicts (3.19) since $d - u \in D$ by assumption (3.18) and $k\beta = \alpha$.

Finally, let us see that $\psi(y_0) \in \operatorname{Str}_{\bar{D}} \psi(F(x_0))$. Indeed, take $y \in F(x_0)$ and assume that $\psi(y) - \psi(y_0) \in -\bar{D}$. Then $\psi(y - y_0) \in -\bar{D}$, and from (3.18) it follows that $y - y_0 \in -D$. As $y_0 \in \operatorname{Str}_D F(x_0)$ by assumption, we conclude that $y - y_0 = 0$, and so $\psi(y) - \psi(y_0) = 0$. □

The next corollary follows immediately from Propositions 3.10 and 3.12.

COROLLARY 3.13. *Let $F : X \to 2^Y$, $x_0 \in S \subset X$, $(x_0, y_0) \in \operatorname{Gr} F$, and $\psi = (\psi_1, \dots, \psi_p) : Y \to \mathbb{R}^p$, with $\psi_i \in D^+$, $i = 1, \dots, p$. In $\mathbb{R}^p$, we consider as ordering cone $\mathbb{R}^p_+$.*

(a) *If $\psi^{-1}(\psi(y_0)) \cap F(x_0) = \{y_0\}$ and $(x_0, \psi(y_0)) \in \operatorname{Str}_{\mathbb{R}^p_+}(\psi \circ F, S; \phi)$, then $(x_0, y_0) \in \operatorname{Str}_D(F, S; \phi)$.*

(b) *If Y is a Banach space, ψ_i, $i = 1, \dots, p$ are linearly independent, $D = \psi^{-1}(\mathbb{R}^p_+)$, and $(x_0, y_0) \in \operatorname{Str}_D(F, S; \phi)$, then $(x_0, \psi(y_0)) \in \operatorname{Str}_{\mathbb{R}^p_+}(\psi \circ F, S; \phi)$.*

The following proposition is proved using some of the ideas developed in Propositions 3.10 and 3.12, and so its proof is omitted.

PROPOSITION 3.14. *Let Y, $\bar{Y}$ be Banach spaces, $\bar{D} \subset \bar{Y}$ a convex cone, $F : X \to 2^Y$, $x_0 \in S \subset X$, and $(x_0, y_0) \in \operatorname{Gr} F$.*

(a) *If $\psi : Y \to \bar{Y}$ is increasing, $\psi^{-1}(\psi(y_0)) \cap F(x_0) = \{y_0\}$ and $(x_0, \psi(y_0)) \in \operatorname{Str}_{\bar{D}}(\psi \circ F, S)$, then $(x_0, y_0) \in \operatorname{Str}_D(F, S)$.*

(b) *If $\psi : Y \to \bar{Y}$ is linear, $D = \psi^{-1}(\bar{D})$ and $(x_0, y_0) \in \operatorname{Str}_D(F, S)$, then $(x_0, \psi(y_0)) \in \operatorname{Str}_{\bar{D}}(\psi \circ F, S)$.*

Note that Propositions 3.10, 3.12, and 3.14 and Corollary 3.13 are also true for the local notions of strict efficiency.

To illustrate the last results we provide an example.

Example 3.15. Let $f = (f_1, \dots, f_p) : X \to \mathbb{R}^p$, $\emptyset \neq I \subset \{1, \dots, p\}$, $\psi : \mathbb{R}^p \to \mathbb{R}^q$ given by $\psi(y_1, \dots, y_p) = (y_i)_{i \in I}$, q being the cardinal of I, $f_I = (f_i)_{i \in I} = \psi \circ f$, and $x_0 \in S \subset X$. Consider in $\mathbb{R}^q$ and $\mathbb{R}^p$ the cones $\mathbb{R}^q_+$ and $D = \psi^{-1}(\mathbb{R}^q_+)$, respectively. Then

(i) $x_0 \in \operatorname{Str}_{\mathbb{R}^q_+}(f_I, S) \Leftrightarrow x_0 \in \operatorname{Str}_D(f, S)$,

(ii) $x_0 \in \operatorname{Str}_{\mathbb{R}^q_+}(f_I, S; \phi) \Leftrightarrow x_0 \in \operatorname{Str}_D(f, S; \phi)$, by Proposition 3.14 and Corollary 3.13, respectively. Moreover, as $\mathbb{R}^p_+ \subset D$, then

(iii) $x_0 \in \mathrm{Str}_{\mathbb{R}^q_+}(f_I, S) \Rightarrow x_0 \in \mathrm{Str}_{\mathbb{R}^p_+}(f, S)$,

(iv) $x_0 \in \mathrm{Str}_{\mathbb{R}^q_+}(f_I, S; \phi) \Rightarrow x_0 \in \mathrm{Str}_{\mathbb{R}^p_+}(f, S; \phi)$.

As a special case, let $f = (f_1, f_2) : X \to \mathbb{R}^2$, where $f_1(x) = \|x\|^m$ and f_2 is an arbitrary function, $x_0 = 0$, $S = X$, $D = \mathbb{R}_+ \times \mathbb{R}$. Then $x_0 \in \mathrm{Str}_D(f, S; m)$ because x_0 is a strict minimizer of order m of f_1.

4. Scalarization.

Scalarization is one of the most important procedures in vector optimization. In this section, we investigate a scalarization process that allows us to transform some notions of strict efficiency for a set-valued map into several notions of minimality for an ordinary scalar function, which is easier to deal with.

Let D be a proper closed convex cone with $\mathrm{int}\, D \neq \emptyset$, and let $e \in \mathrm{int}\, D$ be a fixed point. The Gerstewitz function $g : Y \to \mathbb{R}$ is given by

$$g(y) = \min\{t \in \mathbb{R} : \ te \in y + D\}.$$

It is continuous, convex, increasing, and strictly increasing on Y. This function is well known and widely used in optimization (see, for example, [13, 16, 26]). The following set-to-$\mathbb{R}$ map is introduced in Hernández and Rodríguez-Marín [18] and is an extension of the Gerstewitz function.

Definition 4.1. *The generalized Gerstewitz function $G : 2^Y \to \mathbb{R} \cup \{-\infty\}$ is given by*

$$G(A) = \inf\{t \in \mathbb{R} : \ te \in A + D\}.$$

It is not difficult to check that $G(A) = \inf_{a \in A} g(a)$.

A nonempty set $A \subset Y$ is said to be D-proper if $A + D \neq Y$.

Lemmas 4.2 and 4.4 below can be proved using the ideas of [13]. We give their proofs for reader convenience. A slight variant of the next lemma may be found in [18, Lemma 2.16].

Lemma 4.2. *A is D-proper if and only if $G(A) > -\infty$.*

Proof. We prove only the implication $G(A) = -\infty \Rightarrow A + D = Y$, because the converse is obvious. If $G(A) = -\infty$, then there exists a sequence $t_n \to -\infty$ such that $t_n e \in A + D$ $\forall n$. This implies that $te \in A + D$ $\forall t \in \mathbb{R}$, because if $\alpha e \in A + D$, then $\beta e \in A + D$ $\forall \beta \geq \alpha$ as can be checked. We will check that $Y \subset A + D$. Take any $y \in Y$. Since $e \in \mathrm{int}\, D$, we obtain $e + t^{-1} y \in \mathrm{int}\, D$ for some $t > 0$, and so $y \in -te + \mathrm{int}\, D$. Thus, by the previous remark, $y \in A + D + \mathrm{int}\, D \subset A + D$, which is the desired result. □

Lemma 4.3. (see [9, Lemma 2.5]) *Let A be a nonempty subset of Y. Then*

(i) $\mathrm{int}\,\mathrm{cl}(A + D) = \mathrm{int}(A + D) = A + \mathrm{int}\, D$.

(ii) $\mathrm{cl}(A + D) = \mathrm{cl}(A + \mathrm{int}\, D)$.

The following lemma was proved under the D-properness assumption in [18, Lemma 2.17].

Lemma 4.4. *Let $r \in \mathbb{R}$, and let A be a nonempty subset of Y. The function G has the following properties:*

(i) $G(A) < r \Leftrightarrow re \in A + \mathrm{int}\, D$.

(ii) $G(A) > r \Leftrightarrow re \notin \mathrm{cl}(A + D)$.

(iii) $G(A) = r \Leftrightarrow re \in \mathrm{bd}(A + D)$.

(iv) $G(A) = G(A + D) = G(A + \mathrm{int}\, D) = G(\mathrm{cl}(A + D))$.

Proof. (i) ($\Leftarrow$) As $A + \mathrm{int}\, D$ is an open set, there exists $\varepsilon > 0$ such that $(r - \varepsilon)e = re - \varepsilon e \in A + \mathrm{int}\, D \subset A + D$. Hence, $G(A) \leq r - \varepsilon < r$.

($\Rightarrow$) By the definition of $G(A)$, there exists $t \in \mathbb{R}$, $t < r$, such that $te \in A + D$. Hence

$$re = te + (r - t)e \in A + D + \operatorname{int} D \subset A + \operatorname{int} D.$$

(ii) ($\Leftarrow$) By hypothesis $re \in Y \setminus \operatorname{cl}(A + D)$ and as $Y \setminus \operatorname{cl}(A + D)$ is an open set, there exists $\varepsilon > 0$ such that $(r + \varepsilon)e = re + \varepsilon e \notin \operatorname{cl}(A + D)$. Hence, $(r + \varepsilon)e \notin A + D$, and therefore, $G(A) \geq r + \varepsilon > r$ because if $\alpha e \notin A + D$, then $\beta e \notin A + D$ $\forall \beta \leq \alpha$ as can easily be checked.

($\Rightarrow$) Suppose that $re \in \operatorname{cl}(A + D)$, and let us see that

$$re + \varepsilon e \in A + \operatorname{int} D \quad \forall \varepsilon > 0. \tag{4.1}$$

Indeed, one has that $\operatorname{cl}(A + D) + \operatorname{int} D \subset \operatorname{cl}(A + D)$ as can be verified. Using that $\operatorname{cl}(A+D)+\operatorname{int} D$ is an open set and Lemma (4.3)(i), we derive that $\operatorname{cl}(A+D)+\operatorname{int} D \subset \operatorname{int}\operatorname{cl}(A + D) = A + \operatorname{int} D$. Therefore, (4.1) holds.

Now, from (4.1) it follows that $(r + \varepsilon)e \in A + \operatorname{int} D$ and by part (i), $G(A) < r + \varepsilon$ $\forall \varepsilon > 0$, which contradicts the assumption.

(iii) It follows from parts (i) and (ii), taking into account Lemma (4.3)(i).

(iv) The equality $G(A) = G(A+D)$ is clear from the definition of G since $D+D = D$. In case A is not D-proper, we have $A+D = Y = \operatorname{cl}(A+D)$ and $A+\operatorname{int} D = Y$ by virtue of Lemma 4.3(i), and so part (iv) follows. If A is D-proper, then $G(A) = r \in \mathbb{R}$. One has

$$M := (A + \operatorname{int} D) + D = A + \operatorname{int} D \quad \text{and} \quad N := \operatorname{cl}(A + D) + D = \operatorname{cl}(A + D).$$

By Lemma 4.3, $\operatorname{int}(A + D) = A + \operatorname{int} D$, $\operatorname{cl} M = \operatorname{cl}(A + D)$ and $\operatorname{int} N = A + \operatorname{int} D$. As M is open and N is closed, the sets $A + D$, M, and N have the same closure and the same interior and consequently, the same boundary, i.e.,

$$\operatorname{bd}(A + D) = \operatorname{bd}[(A + \operatorname{int} D) + D] = \operatorname{bd}[\operatorname{cl}(A + D) + D].$$

Now, the conclusion follows from part (iii). □

Remark 4.5. From parts (i) and (ii) of Lemma 4.4, one has the following equivalences:

$$G(A) = 0 \;\Leftrightarrow\; G(A) \leq 0 \text{ and } G(A) \geq 0 \;\Leftrightarrow\; \begin{cases} 0 \in \operatorname{cl}(A + D), \\ 0 \notin A + \operatorname{int} D \text{ (or } A \cap (-\operatorname{int} D) = \emptyset). \end{cases}$$

LEMMA 4.6. *Let $A \subset Y$ and $y_0 \in A$. Then*

$$y_0 \in \operatorname{WMin}_D A \;\Leftrightarrow\; G(A - y_0) = 0.$$

In particular, $y_0 \in \operatorname{Min}_D A$ (or $y_0 \in \operatorname{Str}_D A$) $\Rightarrow$ $G(A - y_0) = 0$.

The proof of this lemma is straightforward from Remark 4.5 (compare with Proposition 2.20 in [18]).

Given the set-valued map $F : X \to 2^Y$ and $(x_0, y_0) \in \operatorname{Gr} F$, we denote $\tilde{F} : X \to 2^Y$ the set-valued map given by

$$\tilde{F}(x) = F(x) - y_0.$$

Next we characterize different kinds of minimizers for a set-valued map F through different kinds of minimizers of the scalar function $G \circ \tilde{F} : X \to \mathbb{R} \cup \{-\infty\}$.

THEOREM 4.7. *Let* $x_0 \in S \subset X$ *and* $(x_0, y_0) \in \operatorname{Gr} F$.

(i) $(x_0, y_0) \in \operatorname{WMin}_D(F, S) \iff x_0 \in \operatorname{Min}(G \circ \tilde{F}, S)$ *and* $G(\tilde{F}(x_0)) = 0$.

(ii) $(x_0, y_0) \in \operatorname{Str}_D(F, S)$ *if* $x_0 \in \operatorname{Str}(G \circ \tilde{F}, S)$ *and* $y_0 \in \operatorname{Str}_D F(x_0)$.

The converse is true if F is D-closed valued, that is, $F(x) + D$ is a closed set for all $x \in \operatorname{Dom} F$.

(iii) $(x_0, y_0) \in \operatorname{Str}_D(F, S; \phi) \iff x_0 \in \operatorname{Str}(G \circ \tilde{F}, S; \phi)$ *and* $y_0 \in \operatorname{Str}_D F(x_0)$.

Proof. (i) By assumption, $(F(x) - y_0) \cap (-\operatorname{int} D) = \emptyset\ \forall x \in S$. So $0 \notin F(x) - y_0 + \operatorname{int} D\ \forall x \in S$ and by Lemma 4.4(i), it follows that $G(F(x) - y_0) \geq 0\ \forall x \in S$. The condition $G(\tilde{F}(x_0)) = 0$ is deduced from Remark 4.5 since $0 \in \operatorname{cl}(F(x_0) - y_0 + D)$ and $0 \notin F(x_0) - y_0 + \operatorname{int} D$. The converse implication follows the inverse path.

(ii) First the "if" part. As $y_0 \in \operatorname{Str}_D F(x_0)$, by Lemma 4.6, $G(F(x_0) - y_0) = 0$. The other hypothesis $x_0 \in \operatorname{Str}(G \circ \tilde{F}, S)$ means that

$$G(F(x) - y_0) > G(F(x_0) - y_0) = 0 \quad \forall x \in S \setminus \{x_0\}.$$

In view of Lemma 4.4(ii), $0 \notin \operatorname{cl}(F(x) - y_0 + D)$. This condition implies that $0 \notin F(x) - y_0 + D$ and then

$$(F(x) - y_0) \cap (-D) = \emptyset \quad \forall x \in S \setminus \{x_0\}.$$

Therefore, $(x_0, y_0) \in \operatorname{Str}_D(F, S)$.

Second, assume that $(x_0, y_0) \in \operatorname{Str}_D(F, S)$ and F is D-closed valued. Then $(F(x) - y_0) \cap (-D) = \emptyset\ \forall x \in S \setminus \{x_0\}$, and so $0 \notin F(x) - y_0 + D = \operatorname{cl}(F(x) - y_0 + D)$. Hence, $G(F(x) - y_0) > 0$ by Lemma 4.4(ii). The condition $0 = G(F(x_0) - y_0)$ follows from Lemma 4.6, since $y_0 \in \operatorname{Str}_D F(x_0)$ by assumption. Therefore, $G(\tilde{F}(x)) > G(\tilde{F}(x_0))\ \forall x \in S \setminus \{x_0\}$.

(iii) ($\Rightarrow$) By assumption there exists $\alpha > 0$ such that

$$(F(x) - y_0 + D) \cap B(0, \alpha\phi(\|x - x_0\|)) = \emptyset \quad \forall x \in S \setminus \{x_0\}.$$

As $B(0, \alpha\phi(\|x - x_0\|))$ is an open set, we have that

$$\operatorname{cl}(F(x) - y_0 + D) \cap B(0, \alpha\phi(\|x - x_0\|)) = \emptyset \quad \forall x \in S \setminus \{x_0\}.$$

This is equivalent to $B(0, \alpha\phi(\|x - x_0\|)) \subset [\operatorname{cl}(F(x) - y_0 + D)]^c$. As $\frac{\alpha}{2\|e\|}\phi(\|x - x_0\|)e \in B(0, \alpha\phi(\|x - x_0\|))$, we derive that $\frac{\alpha}{2\|e\|}\phi(\|x - x_0\|)e \notin \operatorname{cl}(F(x) - y_0 + D)$. Once more using Lemma 4.4(ii), we have that

$$G(F(x) - y_0) > \frac{\alpha}{2\|e\|}\phi(\|x - x_0\|). \tag{4.2}$$

On the other hand, as $(x_0, y_0) \in \operatorname{Str}_D(F, S; \phi)$, also it is satisfied that $y_0 \in \operatorname{Str}_D F(x_0)$. In view of Lemma 4.6, we have $G(F(x_0) - y_0) = 0$. From (4.2), the conclusion follows since $G(\tilde{F}(x)) > G(\tilde{F}(x_0)) + \frac{\alpha}{2\|e\|}\phi(\|x - x_0\|)\ \forall x \in S \setminus \{x_0\}$.

($\Leftarrow$) Assume that there exists $\rho > 0$ such that

$$G\left(\tilde{F}(x)\right) > G\left(\tilde{F}(x_0)\right) + \rho\phi(\|x - x_0\|) \quad \forall x \in S \setminus \{x_0\}.$$

As $y_0 \in \operatorname{Str}_D F(x_0)$, by Lemma 4.6 we have $G(F(x_0) - y_0) = 0$, and so $G(F(x) - y_0) > \rho\phi(\|x - x_0\|)\ \forall x \in S \setminus \{x_0\}$. By Lemma 4.4(ii), $\rho\phi(\|x - x_0\|)e \notin \operatorname{cl}(F(x) - y_0 + D)$, and therefore, $\rho\phi(\|x - x_0\|)e \notin F(x) - y_0 + D$. This implies that

$$[\rho\phi(\|x - x_0\|)e - D] \cap (F(x) - y_0 + D) = \emptyset \quad \forall x \in S \setminus \{x_0\}. \tag{4.3}$$

On the other hand, $0 \in e - \operatorname{int} D$, and so there exists $\delta > 0$ such that $B(0, \delta) \subset e - D$. From here, we derive that $B(0, \delta\rho\phi(\|x - x_0\|)) \subset \rho\phi(\|x - x_0\|)e - D$. Hence, in view of (4.3), it follows that $B(0, \delta\rho\phi(\|x - x_0\|)) \cap (F(x) - y_0 + D) = \emptyset \ \forall x \in S \setminus \{x_0\}$. This implies that $(x_0, y_0) \in \operatorname{Str}_D(F, S; \phi)$ since $y_0 \in \operatorname{Str}_D F(x_0)$. □

If, in particular, F is single-valued, the function G becomes the Gerstewitz function g, and the following result is reached.

Corollary 4.8. *Let $x_0 \in S \subset X$, $f : X \to Y$, and $\tilde{f}(x) = f(x) - f(x_0)$ $\forall x \in X$.*

(i) $x_0 \in \operatorname{WMin}_D(f, S) \iff x_0 \in \operatorname{Min}(g \circ \tilde{f}, S)$.

(ii) $x_0 \in \operatorname{Str}_D(f, S) \iff x_0 \in \operatorname{Str}(g \circ \tilde{f}, S)$.

(iii) $x_0 \in \operatorname{Str}_D(f, S; \phi) \iff x_0 \in \operatorname{Str}(g \circ \tilde{f}, S; \phi)$.

In particular, these results are true for the local notions and also for f the identity, obtaining, in this case, the following corollary. For $y_0 \in Y$, we denote g_{y_0} to the function $g_{y_0}(y) = g(y - y_0)$ $\forall y \in Y$.

Corollary 4.9. *Let $y_0 \in S \subset Y$.*

(i) $y_0 \in \operatorname{WMin}_D S \iff y_0 \in \operatorname{Min}(g_{y_0}, S) \iff g(y - y_0) \geq 0 \ \forall y \in S$.

(ii) $y_0 \in \operatorname{Str}_D S \iff y_0 \in \operatorname{Str}(g_{y_0}, S) \iff g(y - y_0) > 0 \ \forall y \in S \setminus \{y_0\}$.

(iii) $y_0 \in \operatorname{Str}_D(S; \phi) \iff y_0 \in \operatorname{Str}(g_{y_0}, S; \phi)$,

$$y_0 \in \operatorname{Str}_D(S; \phi) \iff \exists \rho > 0 \text{ such that } g(y - y_0) > \rho\phi(\|y - y_0\|) \ \forall y \in S \setminus \{y_0\}.$$

As $\operatorname{Str}_D(S; \phi) \subset \operatorname{Str}_D S \subset \operatorname{Min}_D S$, parts (ii) and (iii) are sufficient conditions for efficiency.

Remark 4.10. If in Theorem 4.7(ii), F is not D-closed valued, the result may be not true. For example, let $A = \{(y_1, y_2) \in \mathbb{R}^2 : \ y_1 < 0, \ y_2 \geq -1/y_1\}$, let $F : \mathbb{R} \to 2^{\mathbb{R}^2}$ be defined as

$$F(x) = \begin{cases} A & \text{if } x \neq 0, \\ \{(0, 0)\} & \text{if } x = 0, \end{cases}$$

$D = \mathbb{R}^2_+$, and $e = (1, 1)$. Then the pair $(x_0, y_0) = (0, (0, 0)) \in \operatorname{Str}_D(F, \mathbb{R})$, but $x_0 \notin \operatorname{Str}(G \circ \tilde{F}, \mathbb{R})$ because $(G \circ \tilde{F})(x) = 0 \ \forall x \in \mathbb{R}$. This example also shows that Proposition 2 in [10] is wrong.

Remark 4.11. Assuming that D is pointed, several kinds of proper efficient points for S are studied in Zaffaroni [39] and are characterized in section 4 (Theorems 4.3, 4.4, and 4.6) through the oriented distance: $\Delta_{-D}(y) = d(y, -D) - d(y, Y \setminus (-D))$. We can reformulate some of these results in terms of ϕ-strict efficiency as follows (for the definition of the new concepts, see [39]):

(i) The following statements are equivalent:
   (a) y_0 is superefficient in S,
   (b) $\exists \alpha > 0$ such that $\Delta_{-D}(y - y_0) > \alpha\|y - y_0\| \ \forall y \in S \setminus \{y_0\}$,
   (c) $y_0 \in \operatorname{Str}(S; 1)$.

   Similar statements are true for the local notions.

(ii) The following statements are equivalent:
   (a) y_0 is strong efficient (we prefer this name instead of strictly efficient used by Zaffaroni to avoid confusion) in S,
   (b) $\exists \phi \in \mathcal{A}$ such that $\Delta_{-D}(y - y_0) > \phi(\|y - y_0\|) \ \forall y \in S \setminus \{y_0\}$,
   (c) $\exists \phi \in \mathcal{A}$ such that $y_0 \in \operatorname{Str}_D(S; \phi)$.

Indeed, we prove only part (ii) because the proof of part (i) is similar (we apply [39, Theorem 4.6] instead of [39, Theorem 4.4]). The equivalence (a) $\Leftrightarrow$ (b) is [39, Theorem 4.4]. Taking into account the relations $\Delta_{-D}(y) > 0 \Leftrightarrow d(y, -D) > 0 \Rightarrow \Delta_{-D}(y) = d(y, -D)$, (b) holds if and only if $\exists \phi \in \mathcal{A}$ such that $d(y - y_0) > \phi(\|y - y_0\|)$

$\forall y \in S \setminus \{y_0\}$, and this implies that $y_0 \in \mathrm{Str}_D(S; \phi)$, i.e., (c) is satisfied. If (c) holds, then $\exists \phi \in \mathcal{A}$, $\exists \alpha > 0$ such that $d(y - y_0) > \alpha\phi(\|y - y_0\|)$ $\forall y \in S \setminus \{y_0\}$. As $\alpha\phi \in \mathcal{A}$, then (b) holds.

According to Corollary 4.9(iii), we can replace in statements (b) the oriented distance Δ_{-D} with the Gerstewitz function g.

For further information on proper efficiency and strict efficiency, see [14, 15].

To finish this section, we provide two examples.

Example 4.12. Let $Y = \mathcal{C}(I)$ be the normed space of all continuous real functions on the compact interval I endowed with the max-norm, and let $D \subset Y$ be the cone of nonnegative functions: $D = \{y \in Y : \ y(t) \geq 0 \ \forall t \in I\}$. Let $e \in \operatorname{int} D$ given by $e(t) = 1 \ \forall t \in I$. It is easy to prove that the Gerstewitz function is $g(y) = \max_{t \in I} y(t)$ $\forall y \in Y$. If we consider $f : \mathbb{R} \to Y$ defined by $f(x)(t) = \exp(x^2 t)$, $x_0 \in \mathbb{R}$, and choose $I = [-1, 1]$, then

$$\left(g \circ \tilde{f}\right)(x) = \max_{t \in [-1,1]} \left\{\exp\left(x^2 t\right) - \exp\left(x_0^2 t\right)\right\} = \begin{cases} \exp\left(x^2\right) - \exp\left(x_0^2\right) & \text{if } |x| \geq |x_0|, \\ \exp\left(-x^2\right) - \exp\left(-x_0^2\right) & \text{if } |x| < |x_0|. \end{cases}$$

If $x_0 \neq 0$, it is clear that $x_0 \in \mathrm{Strl}(g \circ \tilde{f}, \mathbb{R}; 1)$ and so by Corollary 4.8(iii), $x_0 \in \mathrm{Strl}_D(f, \mathbb{R}; 1)$, and if $x_0 = 0$, one has that $x_0 \in \mathrm{Str}(g \circ \tilde{f}, \mathbb{R}; 2)$, and so $x_0 \in \mathrm{Str}_D(f, \mathbb{R}; 2)$.

If now we choose $I = [0, 1]$, then

$$\left(g \circ \tilde{f}\right)(x) = \max_{t \in [0,1]} \left\{\exp\left(x^2 t\right) - \exp\left(x_0^2 t\right)\right\} = \begin{cases} \exp\left(x^2\right) - \exp\left(x_0^2\right) & \text{if } |x| \geq |x_0|, \\ 0 & \text{if } |x| < |x_0|. \end{cases}$$

In this case, if $x_0 \neq 0$, then $x_0 \in \mathrm{Min}(g \circ \tilde{f}, \mathbb{R})$ and so by Corollary 4.8(i), $x_0 \in \mathrm{WMin}_D(f, \mathbb{R})$, and if $x_0 = 0$, one obtains the same as above.

Example 4.13. In vector approximation, one has to approximate not only one but several functions simultaneously. We shall consider a model already discussed in Jahn [20] regarding the following free boundary Stefan problem:

$$u_{xx}(x,t) - u_t(x,t) = 0, \quad (x,t) \in D(s), \tag{4.4}$$

$$u_x(0,t) = h(t), \quad 0 < t \leq T, \tag{4.5}$$

$$u(s(t),t) = 0, \quad 0 < t \leq T, \tag{4.6}$$

$$u_x(s(t),t) = -\dot{s}(t), \quad 0 < t \leq T, \tag{4.7}$$

$$s(0) = 0,$$

where $h \in \mathcal{C}([0,T])$ is a nonpositive function satisfying $h(0) < 0$ and

$$D(s) := \{(x,t) \in \mathbb{R}^2 : \ 0 < x < s(t), \ 0 < t \leq T\} \ \text{ for } s \in \mathcal{C}([0,T])$$

(e.g., see [31, p. 31]). For the approximate solution of this problem, we choose

$$\bar{u}(x,t,a) = \sum_{i=0}^{l} a_i v_i(x,t),$$

with

$$v_i(x,t) = \sum_{k=0}^{[i/2]} \frac{i!}{(i-2k)!k!} x^{i-2k} t^k$$

($[i/2]$ stands for the biggest integer less than or equal to $i/2$) and

$$\bar{s}(t,b) = -h(0)t + \sum_{i=1}^{p} b_i t^{i+1}.$$

For each $a \in \mathbb{R}^{l+1}$, $\bar{u}$ satisfies the partial differential equation (4.4) and for each $b \in \mathbb{R}^p$, $\bar{s}$ satisfies $s(0) = 0$. By substituting $\bar{u}$, $\bar{s}$ into (4.5)–(4.7), we obtain the error functions $\rho_1, \rho_2, \rho_3 \in \mathcal{C}([0,T])$, with

$$\rho_1(t,a,b) := \bar{u}_x(0,t,a) - h(t) = \sum_{i=1,\ i \text{ odd}}^{l} a_i \frac{i!}{((i-1)/2)!} t^{(i-1)/2} - h(t),$$

$$\rho_2(t,a,b) := \bar{u}(\bar{s}(t,b),t,a) = \sum_{i=0}^{l} a_i v_i(\bar{s}(t,b),t),$$

and

$$\rho_3(t,a,b) := \bar{u}_x(\bar{s}(t,b),t,a) + \dot{\bar{s}}(t,b) = \sum_{i=1}^{l} a_i v_{ix}(\bar{s}(t,b),t) + \dot{\bar{s}}(t,b).$$

If $||\cdot||$ is any norm in $\mathcal{C}([0,T])$, we formulate the following vector optimization problem for the approximate solution of the free boundary Stefan problem: determine minimal or weakly minimal elements of the set

$$\left\{(||\rho_1(\cdot,a,b)||, ||\rho_2(\cdot,a,b)||, ||\rho_3(\cdot,a,b)||) \in \mathbb{R}^3 :\ (a,b) \in \mathbb{R}^{l+1} \times \mathbb{R}^p\right\},$$

where the vector space $\mathbb{R}^3$ is ordered by the usual nonnegative orthant $\mathbb{R}^3_+$.

By setting $f_i(a,b) = ||\rho_i(\cdot,a,b)||$, $(a,b) \in \mathbb{R}^{l+1} \times \mathbb{R}^p$, and $f = (f_1, f_2, f_3)$, we are concerned with the scalar function $g \circ \tilde{f}$, where $\tilde{f}(a,b) = f(a,b) - f(\bar{a},\bar{b})$ and $g(y) = \max_{1 \leq i \leq 3} y_i$. The function g was obtained by taking $e = (1,1,1)$. By applying Corollary 4.8, we obtain the following result.

Corollary 4.14. *Consider the above free boundary Stefan problem, and let $(\bar{a},\bar{b}) \in \mathbb{R}^{l+1} \times \mathbb{R}^p$. Then*

(a) *$(\bar{a},\bar{b})$ is a weak minimizer if and only if* $\max_{i=1,2,3}\{f_i(a,b) - f_i(\bar{a},\bar{b})\} \geq 0$ $\forall (a,b) \in \mathbb{R}^{l+1} \times \mathbb{R}^p$.

(b) *$(\bar{a},\bar{b})$ is a strict minimizer if and only if* $\max_{i=1,2,3}\{f_i(a,b) - f_i(\bar{a},\bar{b})\} > 0$ $\forall (a,b) \in \mathbb{R}^{l+1} \times \mathbb{R}^p$*, with* $(a,b) \neq (\bar{a},\bar{b})$.

5. Optimality conditions for strict minimizers.

In this section, we focus on minimizers of order one. We provide several optimality conditions through different kind of derivatives. In general, a derivative is a local notion, so the minimizers connected are local, but we also provide some results with global derivatives for global minimizers. Optimality conditions (specially necessary conditions) is a topic widely treated in the literature. See [7, 10, 11, 12, 21, 27, 36, 37] as examples in a similar framework as the one considered here.

Definition 5.1. *Let* $(x_0,y_0) \in \operatorname{Gr} F$.

(a) *The radial derivative of F at (x_0,y_0) (see* [12, 37]*) is the set-valued map $D_R F(x_0,y_0) : X \to 2^Y$ defined by $u \in D_R F(x_0,y_0)(v)$ if and only if*

$$(5.1)\ \exists\, t_n > 0,\ \exists\, v_n \to v,\ \exists\, u_n \to u \text{ such that } y_n := y_0 + t_n u_n \in F(x_0 + t_n v_n)\ \ \forall\, n.$$

(b) *(Aubin* [2]*.) The contingent derivative of* F *at* (x_0, y_0) *is the set-valued map* $D_cF(x_0, y_0) : X \to 2^Y$ *defined by* $u \in D_cF(x_0, y_0)(v)$ *if and only if*

$$(5.2)\ \exists\, t_n \to 0^+,\ \exists\, v_n \to v,\ \exists\, u_n \to u \text{ such that } y_n := y_0 + t_n u_n \in F(x_0 + t_n v_n)\ \ \forall\, n.$$

(c) *(Shi* [34, *Definition* 2.4]*.) The Shi derivative of* F *at* (x_0, y_0) *is the set-valued map* $D_SF(x_0, y_0) : X \to 2^Y$ *defined by* $u \in D_SF(x_0, y_0)(v)$ *if and only if*

$$(5.3)\quad \exists\, t_n > 0,\ \exists\, v_n \to v,\ \exists\, u_n \to u \text{ such that } t_n v_n \to 0 \text{ and } \\ y_n := y_0 + t_n u_n \in F(x_0 + t_n v_n)\ \ \forall\, n.$$

Condition (5.1) is equivalent to $(v, u) \in \operatorname{cl}\operatorname{cone}(\operatorname{Gr} F - (x_0, y_0))$, so $\operatorname{Gr} D_RF(x_0, y_0) = \operatorname{cl}\operatorname{cone}(\operatorname{Gr} F - (x_0, y_0))$; condition (5.2) is equivalent to $(v, u) \in T(\operatorname{Gr} F, (x_0, y_0))$.

Some properties of the previous derivatives are collected in the following lemma. Recall that a set-valued map F is positively homogeneous if $F(tx) = tF(x)$ $\forall t > 0$, $x \in \operatorname{Dom} F$.

LEMMA 5.2. (i) *The set-valued maps* $D_RF(x_0, y_0)$*,* $D_cF(x_0, y_0)$*, and* $D_SF(x_0, y_0)$ *are positively homogeneous.*

(ii) $F(x) - y_0 \subset D_RF(x_0, y_0)(x - x_0)$ $\forall\, x \in X$*.*

(iii) $D_cF(x_0, y_0)(v) \subset D_SF(x_0, y_0)(v) \subset D_RF(x_0, y_0)(v)$ $\forall\, v \in X$*.*

(iv) $D_cF(x_0, y_0)(v) = D_SF(x_0, y_0)(v)$ $\forall\, v \in X \setminus \{0\}$*.*

Parts (i)–(iii) can be found in Taa [37, Remarks 2.3 and 2.5 and Theorem 2.1]; part (iv) follows easily from the definitions.

DEFINITION 5.3. *A set-valued map* F *is said to be compact at* x_0 [29] *if for any sequence* $(x_n, y_n) \in \operatorname{Gr} F$ *such that* $x_n \to x_0$*, there exists a convergent subsequence* $(x_{n_k}, y_{n_k}) \to (x_0, y)$ *for some* $y \in F(x_0)$*. Whenever this is true for each* $x_0 \in A \subset \operatorname{Dom} F$*, we say that* F *is compact on the set* A*.*

We point out that if F is compact on A, then F is compact-valued on A ($F(x)$ is compact for all $x \in A$). Indeed, take $x \in A$ and a sequence $y_n \in F(x)$. Then choosing $x_n = x$, we have that $(x_n, y_n) \in \operatorname{Gr} F$, with $x_n \to x$, and as F is compact, there exists a convergent subsequence (y_{n_k}) to some $y \in F(x)$.

We start by characterizing a (global) minimizer of order one for an unconstrained problem in terms of the radial derivative.

THEOREM 5.4. *Let* $F : X \to 2^Y$ *and* $(x_0, y_0) \in \operatorname{Gr} F$*. If* $(x_0, y_0) \in \operatorname{Str}_D(F, X; 1)$*, then* $D_RF(x_0, y_0)(v) \cap (-D) = \emptyset$ $\forall\, v \in X \setminus \{0\}$*. The converse is true if* D *is closed,* X *is finite-dimensional,* $y_0 \in \operatorname{Str}_D F(x_0)$*, and the set-valued map* $D_RF(x_0, y_0)$ *is compact on the set* $S^1 := \{v \in X :\ \|v\| = 1\}$*.*

Proof. ($\Rightarrow$) By hypothesis, there exists $\alpha > 0$ such that

$$(F(x) + D) \cap B(y_0, \alpha\|x - x_0\|) = \emptyset \quad \forall\, x \in X \setminus \{x_0\},$$

i.e., $\frac{F(x) - y_0}{\|x - x_0\|} \subset (B(0, \alpha) - D)^c$. Assume that there is $u \in D_RF(x_0, y_0)(v) \cap (-D)$ for some $v \in X \setminus \{0\}$. Then (5.1) is fulfilled, and so $\frac{y_n - y_0}{t_n} = u_n \to u \in -D$. Denoting $x_n = x_0 + t_n v_n$, we have $\frac{x_n - x_0}{t_n} = v_n \to v$ and

$$\frac{y_n - y_0}{\|x_n - x_0\|} = \frac{t_n}{\|x_n - x_0\|} \cdot \frac{y_n - y_0}{t_n} = \frac{1}{\|v_n\|} u_n \to \frac{1}{\|v\|} u \in -D.$$

On the other hand,

$$\frac{y_n - y_0}{\|x_n - x_0\|} \in \frac{F(x_n) - y_0}{\|x_n - x_0\|} \subset (B(0, \alpha) - D)^c.$$

Since the latter set is closed, we obtain that $\frac{1}{\|v\|}u \in (B(0,\alpha) - D)^c \cap (-D)$, yielding a contradiction since $(B(0,\alpha) - D)^c \cap (-D) = \emptyset$. This last equality holds because $-D \subset (-D) + B(0,\alpha)$. Let us observe that $v \neq 0$ implies that $x_n - x_0 \neq 0$ for all n sufficiently large.

($\Leftarrow$) To simplify, let us denote $H(v) = D_R F(x_0, y_0)(v)$. Let $\varphi : S^1 \to \mathbb{R}$ be defined by

$$\varphi(v) = \inf\{d(u, -D) : u \in H(v)\}. \tag{5.4}$$

This infimum is attained because $H(v)$ is compact and the function $Y \ni u \mapsto d(u,-D) \in \mathbb{R}$ is Lipschitz [3] (in particular, it is continuous). Hence, there exists $u \in H(v)$ such that $d(u,-D) = d(H(v),-D) = \varphi(v)$. Moreover, $d(u,-D) > 0$, since otherwise $u \in -D$, contradicting the hypothesis.

Now, let $\alpha = \inf\{\varphi(v) :\ v \in S^1\}$, and let us see that $\alpha > 0$. Suppose that $\alpha = 0$, then for each $n \in \mathbb{N}$ there exist $v_n \in S^1$ and $u_n \in H(v_n)$ such that

$$\varphi(v_n) = d(u_n, -D) < \frac{1}{n}. \tag{5.5}$$

As S^1 is compact, we can assume, taking a subsequence if necessary, that $v_n \to v_0$ for some $v_0 \in S^1$. As H is compact on S^1, there exists a subsequence $u_{n_k} \in H(v_{n_k})$ converging to some $u_0 \in H(v_0)$. As $d(\cdot, -D)$ is continuous, from (5.5) it follows that $d(u_0,-D) = 0$. But $u_0 \in H(v_0)$ implies that $0 < \varphi(v_0) \leq d(u_0,-D)$, and we have a contradiction. Thus,

$$\inf\{d(D_R F(x_0,y_0)(v), -D) :\ \|v\| = 1\} = \alpha > 0.$$

As $D_R F(x_0,y_0)$ is positively homogeneous and $d(D_R F(x_0,y_0)(v/\|v\|), -D) \geq \alpha\ \forall\, v \in X \setminus \{0\}$, we deduce that $d(D_R F(x_0,y_0)(v), -D) \geq \alpha\|v\|\ \forall\, v \in X$. Now by Lemma 5.2(ii), $F(x_0+v) - y_0 \subset D_R F(x_0,y_0)(v)$, and therefore,

$$d(F(x_0+v) - y_0, -D) \geq d(D_R F(x_0,y_0)(v), -D) \geq \alpha\|v\|.$$

In other terms, $d(F(x) - y_0, -D) \geq \alpha\|x - x_0\|\ \forall\, x \in X \setminus \{x_0\}$. According to (3.3), this proves that $(x_0,y_0) \in \mathrm{Str}_D(F, X; 1)$. □

The second part of Theorem 5.4 improves the conclusion of Theorem 3.1(i) in [37].

Next we establish our basic assumption (B) (in terms of the contingent derivative of the set-valued map $F+D$) implying local optimality. Afterwards, we provide several sufficient conditions related to those existing in the literature but stronger than the basic assumption.

Assumption (B).

$(B1)$ $0 \notin D_c(F + D)(x_0,y_0)(v)\ \ \forall\, v \in T(S, x_0) \setminus \{0\}$.

$(B2)$ $y_0 \in \mathrm{Str}_D\, F(x_0)$.

THEOREM 5.5. *Let $F : X \to 2^Y$, $x_0 \in S \subset X$, and $(x_0,y_0) \in \mathrm{Gr}\, F$. Suppose that X is finite-dimensional and that Assumption (B) holds, then $(x_0,y_0) \in$ $\mathrm{Strl}_D(F, S; 1)$.*

Proof. Assume that the conclusion is false. Then, by Proposition 3.4, there exist sequences $x_n \in S \setminus \{x_0\}$, $y_n \in F(x_n)$, $d_n \in D$ such that $x_n \to x_0$ and

$$\lim_{n\to\infty} \frac{y_n - y_0 + d_n}{\|x_n - x_0\|} = 0. \tag{5.6}$$

As X is finite-dimensional, we can suppose that

$$v_n := \frac{x_n - x_0}{\|x_n - x_0\|} \to v \in T(S, x_0) \setminus \{0\}. \tag{5.7}$$

It turns out that $y_n + d_n - y_0 \in (F + D)(x_n) - y_0$, so (5.6) and (5.7) imply that $0 \in D_c(F + D)(x_0, y_0)(v)$, which contradicts $(B1)$. □

Although the proof of the previous theorem is somewhat simple, it contains several important results established elsewhere, the proofs of which are rather involved.

Theorem 5.5 improves the first part of Theorem 3.2 in Durea [11] in the following senses: the convex cone D in [11] must be closed and pointed, and the final space Y must be finite-dimensional. We do not require these conditions. Also, our conclusion is sharper. Moreover, in [11] a special space $\overline{Y}$ with a special topology and an extended notion of contingent derivative are considered.

For a real function $h : X \to \mathbb{R}$, the lower Hadamard derivative of h at x_0 in the direction v is

$$\underline{d}h(x_0)(v) = \liminf_{(t,v') \to (0^+, v)} t^{-1}(h(x_0 + tv') - h(x_0)).$$

It is not difficult to verify that

$$D_c(h + \mathbb{R}_+)(x_0, h(x_0))(v) \subset [\underline{d}h(x_0)(v), +\infty), \tag{5.8}$$

since if $u = \lim_n t_n^{-1}(h(x_0 + t_n v_n) + \alpha_n - h(x_0))$, with $(t_n, v_n) \to (0^+, v)$ and $\alpha_n \in \mathbb{R}_+$, then $\underline{d}h(x_0)(v) \le \liminf_n t_n^{-1}(h(x_0 + t_n v_n) - h(x_0)) \le u$.

Corollary 5.6. *Let $f = (f_1, \dots, f_p) : X \to \mathbb{R}^p$ and $x_0 \in S \subset X$. If X is finite-dimensional and*

$$\forall\, v \in T(S, x_0) \setminus \{0\}\ \exists i = 1, \dots, p \text{ such that } \underline{d}f_i(x_0)(v) > 0,$$

then $x_0 \in \operatorname{Strl}_{\mathbb{R}^p_+}(f, S; 1)$.

Proof. One has

$$D_c(f + \mathbb{R}^p_+)(x_0, f(x_0))(v) \subset \prod_{i=1}^{p} D_c(f_i + \mathbb{R}_+)(x_0, f_i(x_0))(v) \subset \prod_{i=1}^{p} [\underline{d}f_i(x_0)(v), +\infty). \tag{5.9}$$

Indeed, the second inclusion follows from (5.8). For the first one, take $u = (u_1, \dots, u_p) \in D_c(f + \mathbb{R}^p_+)(x_0, f(x_0))(v)$, then $t_n^{-1}(f(x_0 + t_n v_n) + d_n - f(x_0)) \to u$, with $(t_n, v_n) \to (0^+, v)$ and $d_n = (d_{n,1}, \dots, d_{n,p}) \in \mathbb{R}^p_+$. Hence, $t_n^{-1}(f_i(x_0 + t_n v_n) + d_{n,i} - f_i(x_0)) \to u_i$ $\forall\, i = 1, \dots, p$ and consequently, $u_i \in D_c(f_i + \mathbb{R}_+)(x_0, f_i(x_0))(v)$.

If $(B1)$ is false for $F = f$, then $0 \in D_c(f + \mathbb{R}^p_+)(x_0, f(x_0))(v)$ for some $v \in T(S, x_0) \setminus \{0\}$. In view of (5.9), we have a contradiction to the hypothesis. So $(B1)$ holds, and the conclusion follows by applying Theorem 5.5. □

The conclusion of Corollary 5.6 is somewhat weaker than Corollary 3.1(ii) in [23], but its proof is easier. Corollary 5.6 improves the conclusion of Theorem 2.2 in [8], and if $p = 1$, it also improves the conclusions of Corollary 3.2 in [11] (which is the constrained version of Theorem 2.1 in [30]) and Theorem 2.3(i) in [19].

In order to provide some conditions implying the validity of assumption (B), let us give a preliminary result and some definitions.

Lemma 5.7. *Let $F : X \to 2^Y$, $(x_0, y_0) \in \operatorname{Gr} F$, and $v \in X \setminus \{0\}$. Then*

$$0 \notin D_c(F + D)(x_0, y_0)(v) \iff D_c(F + D)(x_0, y_0)(v) \cap (-D) = \emptyset.$$

Proof. ($\Leftarrow$) It is obvious.

($\Rightarrow$) Assume that $-d \in D_c(F+D)(x_0,y_0)(v) \cap (-D)$. Then there exist $t_n \to 0^+$, $v_n \to v$, $d_n \in D$, and $y_n \in F(x_n)$, with $x_n = x_0 + t_n v_n$ such that $\frac{y_n+d_n-y_0}{t_n} \to -d$. Thus, $\frac{y_n+d_n+t_n d-y_0}{t_n} \to 0$. Since D is a convex cone, the previous limit says that $0 \in D_c(F+D)(x_0,y_0)(v)$, which is a contradiction. □

We have required that $v \neq 0$ in the preceding lemma because $0 \in D_c(F+D)(x_0,y_0)(0)$.

Definition 5.8. (a) [1] *A set-valued map F is said to be compactly approximable at $(x_0,y_0) \in \operatorname{Gr} F$ if, for each $v \in X$, there exist a set-valued map R from X into the set of all nonempty compact subsets of Y, a neighborhood V of v, and a function $r : (0,1) \times V \to (0,+\infty)$ satisfying*

(i) $\lim_{(t,u)\to(0^+,v)} r(t,u) = 0$,

(ii) $\forall (t,u) \in (0,1) \times V$, *one has* $F(x_0+tu) \subset y_0 + t(R(v) + r(t,u)\operatorname{cl} B(0,1))$.

(b) [7, *Definition* 7] *A set-valued map F is said to be directionally compact at $(x_0,y_0) \in \operatorname{Gr} F$ in the direction $v \in X$ if $\forall t_n \to 0^+$, $\forall v_n \to v$, $\forall u_n \in Y$, with $y_0 + t_n u_n \in F(x_0+t_n v_n)$ $\forall n$, there exists a subsequence (u_{n_k}) converging to some $u \in Y$.*

If F is single-valued and Fréchet differentiable at x_0, then F is directionally compact at (x_0,y_0) in any direction $v \in X$.

In connection to Assumption (B), we have two sets of conditions:

Assumption (A).

$(A1)$ $D_cF(x_0,y_0)(v) \cap (-D) = \emptyset \ \forall v \in T(S,x_0) \setminus \{0\}$.

$(A2)$ F is directionally compact at (x_0,y_0) in each direction $v \in T(S,x_0) \setminus \{0\}$.

$(A3)$ $y_0 \in \operatorname{Str}_D F(x_0)$.

Assumption (S).

$(S1)$ $D_cF(x_0,y_0)(v) \cap (-D) = \emptyset \ \forall v \in T(S,x_0) \setminus \{0\}$.

$(S2)$ $D_SF(x_0,y_0)(0) \cap (-D) = \{0\}$.

$(S3)$ D has a compact base.

Theorem 5.9. *The following assertions hold:*

(a) *Assumption* $(S) \Rightarrow$ *assumption* (B),

(b) *Assumption* $(A) \Rightarrow$ *assumption* (B).

Consequently, if X is finite-dimensional, each of the assumptions (S) and (A) ensure that $(x_0,y_0) \in \operatorname{Strl}_D(F,S;1)$.

Proof. $(S) \Rightarrow (B)$. According to Proposition 3.1 in Shi [34] (his result is still valid without the pointedness of D), $(S2)$ and $(S3)$ imply the following condition:

$$(BC) \quad D_c(F+D)(x_0,y_0)(v) = D_cF(x_0,y_0)(v) + D \quad \forall v \in X.$$

Now, let us prove that

$$(5.10) \qquad (S1) + (BC) \ \Rightarrow \ (B1).$$

Indeed, $(S1)$ is equivalent to

$$[D_cF(x_0,y_0)(v) + D] \cap (-D) = \emptyset \quad \forall v \in T(S,x_0) \setminus \{0\}$$

as can be easily checked. Using (BC), this is equivalent to

$$D_c(F+D)(x_0,y_0)(v) \cap (-D) = \emptyset \quad \forall v \in T(S,x_0) \setminus \{0\}$$

and by Lemma 5.7, it is equivalent to condition $(B1)$.

We need only to prove condition $(B2)$. Assume that $y_0 \notin \operatorname{Str}_D F(x_0)$, then there is $y \in F(x_0)$ such that $y - y_0 \in -D \setminus \{0\}$. Choose the following sequences: $t_n = 1$, $v_n = 0$, and $y_n = y \in F(x_0 + t_n v_n)$ $\forall n$. Then $v_n \to 0$, $t_n v_n \to 0$, and $\frac{y_n - y_0}{t_n} \to y - y_0$, so $y - y_0 \in D_S F(x_0, y_0)(0) \cap (-D)$, which contradicts $(S2)$.

$(A) \Rightarrow (B)$. By Proposition 5 in [7], $(A2)$ implies (BC). Now since $(A1) = (S1)$, from (5.10) we have $(B1)$, and as $(A) = (B2)$, the proof is completed. □

Remark 5.10. (1) A condition which is more easy verifiable than $(A2)$ is the following:

$(A2')$ F is compactly approximable at (x_0, y_0).

More precisely, we have that $(A2')$ implies $(A2)$. Indeed, choose sequences (t_n), (v_n), and (u_n) satisfying Definition 5.8(b), and let $y_n = y_0 + t_n u_n$. By assumption, there exist sequences $w_n \in R(v)$, $r_n > 0$, and $b_n \in \operatorname{cl} B(0,1)$ such that

$$\frac{y_n - y_0}{t_n} = w_n + r_n b_n \quad \text{and} \quad r_n \to 0^+.$$

As $R(v)$ is a compact set of Y, (w_n) possesses a convergent subsequence $w_{n_k} \to w \in Y$. Hence,

$$u_{n_k} = \frac{y_{n_k} - y_0}{t_{n_k}} = w_{n_k} + r_{n_k} b_{n_k} \to w,$$

since $r_{n_k} b_{n_k} \to 0$.

(2) If F is compact at $(x_0, y_0) \in \operatorname{Gr} F$, $y_0 \in \operatorname{Str}_D F(x_0)$, and $D_c F(x_0, y_0)(0) \cap (-D) = \{0\}$, then $(S2)$ holds. This follows from Theorem 2.1 in Taa [37]. We require $y_0 \in \operatorname{Str}_D F(x_0)$ instead of $y_0 \in \operatorname{Min}_D F(x_0)$ so that the proof of this author is valid for a nonpointed cone D.

(3) Condition (BC) plays a crucial role in the proof of the theorem. We refer to Bednarczuk and Song [7] for other conditions implying (BC).

By taking into account Lemma 5.2(iii), Theorem 5.9, under Assumption (S), generalizes Theorem 4.1 in Taa [36] established for an unconstrained problem $(S = X)$.

If $f : X \to Y$ is Hadamard differentiable at x_0, i.e., for any $v \in X$ there exists $df(x_0)(v) = \lim_{t \to 0^+,\, v' \to v} t^{-1}(f(x_0 + tu) - f(x_0))$, then Theorem 5.9, under Assumption (A), collapses into the sufficient conditions of Theorem 4.1 in [24], because $D_c f(x_0, f(x_0))(v) = \{df(x_0)(v)\}$ and f is obviously compactly approximable at $(x_0, f(x_0))$.

In order to obtain more verifiable conditions than $(S2)$, we introduce the notion of D-calmness for a set-valued map.

Definition 5.11. *A set-valued map* $F : X \to 2^Y$ *is said to be* D*-calm at* $(x_0, y_0) \in \operatorname{Gr} F$ *if there exist a neighborhood* U *of* x_0 *and a constant* $L > 0$ *such that*

$$F(x) \subset y_0 + D + L\|x - x_0\| \operatorname{cl} B(0,1) \quad \forall\, x \in U. \tag{5.11}$$

If F is single-valued and $D = \{0\}$, this concept is the usual notion of calmness [32, section F, Chapter 8] also named stable function (see, e.g., [25]).

Lemma 5.12. *If* D *is closed and* F *is* D*-calm at* $(x_0, y_0) \in \operatorname{Gr} F$*, then*

$$D_S F(x_0, y_0)(0) \subset D.$$

Proof. Take $u \in D_S F(x_0, y_0)(0)$, then there exist sequences $t_n > 0$, $v_n \to 0$, and $u_n \to u$ such that $t_n v_n \to 0$ and $y_n := y_0 + t_n u_n \in F(x_0 + t_n v_n)$. Hence,

$\frac{y_n-y_0}{t_n}=u_n\to u$ and for n large enough, $x_n:=x_0+t_nv_n\in U$, where U satisfies (5.11). Therefore, there exist $d_n\in D$ and $b_n\in\operatorname{cl}B(0,1)$ such that $y_n=y_0+d_n+L\|x_n-x_0\|b_n$. Consequently,

$$\frac{y_n-y_0}{t_n}=\frac{d_n}{t_n}+L\|v_n\|b_n,$$

and as $L\|v_n\|b_n\to 0$, we have that $u=\lim t_n^{-1}d_n\in D$ since D is a closed cone. □

COROLLARY 5.13. *Let $F:X\to 2^Y$ and $x_0\in S\subset X$. Suppose that X is finite-dimensional, D has a compact base and is pointed, and F is D-calm at $(x_0,y_0)\in\operatorname{Gr}F$. If*

$$D_cF(x_0,y_0)(v)\cap(-D)=\emptyset\quad\forall\, v\in T(S,x_0)\setminus\{0\},$$

then $(x_0,y_0)\in\operatorname{Strl}_D(F,S;1)$.

Proof. As D has a compact base, D is closed. By Lemma 5.12, we see that $D_SF(x_0,y_0)(0)\subset D$, and as D is pointed, we have that condition $(S2)$ is satisfied. Now the result follows from Theorem 5.9. □

Corollary 5.13 is a generalization of Theorem 5.9 in [25] and Theorem 3.3 in [8], which are valid for stable functions $f:X\to Y$, but this class of functions is D-calm.

Now, we present a result that is very close to the converse of Theorem 5.5. To that purpose, we recall that the cone of attainable directions to S at $x_0\in\operatorname{cl}S$ is

$$A(S,x_0)=\{v\in X:\ \forall\, t_n\to 0^+,\ \exists\, v_n\to v \text{ such that } x_0+t_nv_n\in S\ \forall\, n\in\mathbb{N}\}.$$

It is said that S is derivable at $x_0\in S$ if $T(S,x_0)=A(S,x_0)$. We also need a new notion.

DEFINITION 5.14. *A set-valued map $F:X\to 2^Y$ is said to be D-pseudo Lipschitz at $(x_0,y_0)\in\operatorname{Gr}F$ if there exist neighborhoods U of x_0 and V of y_0 and a constant $L>0$ such that*

$$F(x)\cap V\subset F(x')+D+L\|x-x'\|\operatorname{cl}B(0,1)\quad\forall\, x,x'\in U. \tag{5.12}$$

This notion is new in our knowledge and weaker than other similar ones, as the following relations and example show:

(i) F Lipschitz ⇒ F pseudo Lipschitz ⇒ F D-pseudo Lipschitz.

(ii) F Lipschitz ⇒ F D-Lipschitz ⇒ $F+D$ D-Lipschitz ⇒ $F+D$ D-pseudo Lipschitz ⇒ F D-pseudo Lipschitz.

Recall that F is pseudo Lipschitz if (5.12) holds with $D=\{0\}$, F is D-Lipschitz [10] if (5.12) holds with $V=Y$, and F is Lipschitz if (5.12) holds with $V=Y$ and $D=\{0\}$.

Example 5.15. (a) Let $f:\mathbb{R}\to\mathbb{R}^2$ be given by $f(x)=(x,\sqrt[3]{x})$, $x_0=0$, $y_0=(0,0)$, and $D=\mathbb{R}_+\times\mathbb{R}$. Then f is D-pseudo Lipschitz at (x_0,y_0), but f is not pseudo Lipschitz at (x_0,y_0). To check the first assertion, it is enough to observe that $d(f(x),f(x')+D)=\max\{x'-x,0\}\le|x-x'|$. For the second assertion, choose a sequence $t_n\to 0^+$, and let $s_n=t_n/2$. Then

$$w_n:=\frac{f(t_n)-f(s_n)}{t_n-s_n}=\left(1,\frac{1}{3\sqrt[3]{c_n^2}}\right),$$

for some $c_n\in(s_n,t_n)$, and therefore, for any $L>0$, $w_n\notin L\operatorname{cl}B(0,1)$ for all n large enough.

(b) Let $F : \mathbb{R} \to 2^{\mathbb{R}}$ be given by

$$F(x) = \begin{cases} \{0\} & \text{if } x \neq 0, \\ \{-1, 0\} & \text{if } x = 0, \end{cases}$$

$x_0 = 0$, $y_0 = 0$, and $D = \mathbb{R}_+$. Then F is D-pseudo Lipschitz at (x_0, y_0), but $F + D$ is not D-pseudo Lipschitz at (x_0, y_0) as can be verified.

There is no relationship between D-calmness and D-pseudo Lipschitzianity: the set-valued map F of Example 5.15(b) is D-pseudo Lipschitz at $(0, 0)$, but it is not D-calm at $(0, 0)$; F is D-calm at $(0, -1)$, but it is not D-pseudo Lipschitz at $(0, -1)$. However, if f is a function, f D-pseudo Lipschitz at $(x_0, f(x_0))$ implies f is D-calm at $(x_0, f(x_0))$.

THEOREM 5.16. *Let $F : X \to 2^Y$, $x_0 \in S \subset X$, and $(x_0, y_0) \in \operatorname{Gr} F$. If F is D-pseudo Lipschitz at (x_0, y_0) and $(x_0, y_0) \in \operatorname{Strl}_D(F, S; 1)$, then*

$$D_cF(x_0, y_0)(v) \cap (-D) = \emptyset \quad \forall\, v \in A(S, x_0) \setminus \{0\}. \tag{5.13}$$

Proof. F is D-pseudo Lipschitz at (x_0, y_0) means that there exist neighborhoods U of x_0 and V of y_0 and a constant $L > 0$ such that (5.12) holds. Assume that there is $u \in D_cF(x_0, y_0)(v) \cap (-D)$ for some $v \in A(S, x_0) \setminus \{0\}$. Then there exist $t_n \to 0^+$, $v_n \to v$, and $y_n \in F(x_0 + t_n v_n)$ such that $u_n = \frac{y_n - y_0}{t_n} \to u$. From here,

$$\frac{y_n - y_0 - t_n u}{t_n} \to 0. \tag{5.14}$$

Let us observe that $y_n \to y_0$ and putting $x_n = x_0 + t_n v_n \in X$, we have $x_n \to x_0$.

As $v \in A(S, x_0)$, for the sequence (t_n), there exists $v'_n \to v$ such that $x'_n := x_0 + t_n v'_n \in S$. From these expressions, it results that

$$\frac{\|x'_n - x_0\|}{t_n} \to \|v\|. \tag{5.15}$$

As $y_n \in F(x_n) \cap V$, $x_n, x'_n \in U$ for all n large enough, it follows from (5.12) that there exist $y'_n \in F(x'_n)$, $d_n \in D$, $b_n \in \operatorname{cl} B(0, 1)$ such that $y_n = y'_n + d_n + L t_n \|v_n - v'_n\| b_n$. Now,

$$\begin{aligned} \frac{t_n}{\|x'_n - x_0\|} \cdot \frac{y_n - y_0 - t_n u}{t_n} &= \frac{y'_n + d_n + L t_n \|v_n - v'_n\| b_n - y_0 - t_n u}{\|x'_n - x_0\|} \\ &= \frac{y'_n - y_0 + d_n - t_n u}{\|x'_n - x_0\|} + L \|v_n - v'_n\| \frac{t_n}{\|x'_n - x_0\|} b_n. \end{aligned}$$

From here, taking into account (5.14), (5.15), and that $\|v_n - v'_n\| \to 0$, it follows that $\frac{y'_n - y_0 + d_n - t_n u}{\|x'_n - x_0\|} \to 0$, with $y'_n \in F(x'_n)$, $d_n - t_n u \in D$, $x'_n \to x_0$, and $x'_n \in S$ and by applying Proposition 3.4 with $\phi(t) = t$, we conclude that $(x_0, y_0) \notin \operatorname{Strl}_D(F, S; 1)$, which is a contradiction. □

We present a characterization for strict minimizers of order one under very general conditions.

COROLLARY 5.17. *Let $F : X \to 2^Y$, $x_0 \in S \subset X$, and $(x_0, y_0) \in \operatorname{Gr} F$. Assume that D is pointed, X is finite-dimensional, $F + D$ is D-pseudo Lipschitz at (x_0, y_0), and S is derivable at x_0. Then, $(x_0, y_0) \in \operatorname{Strl}_D(F, S; 1)$ if and only if Assumption (B) is satisfied.*

Proof. The "if" part is clear from Theorem 5.5. Let us see the "only if" part. Condition $(B2)$ is obvious from the definition of $\mathrm{Strl}_D(F, S; 1)$. By hypothesis $(x_0, y_0) \in \mathrm{Strl}_D(F, S; 1)$ and by Proposition 3.3, $(x_0, y_0) \in \mathrm{Strl}_D(F + D, S; 1)$. By taking into account that $A(S, x_0) = T(S, x_0)$, Theorem 5.16 applied to $F + D$ instead of F implies that

$$D_c(F + D)(x_0, y_0)(v) \cap (-D) = \emptyset \quad \forall\, v \in T(S, x_0) \setminus \{0\}.$$

In view of Lemma 5.7, this last condition is equivalent to $(B1)$. □

To illustrate the above results, we give some examples.

Example 5.18. Let l^2 be the Hilbert space of sequences of real numbers (a_i) such that $a_1^2 + a_2^2 + \cdots < \infty$, and let $f : \mathbb{R} \to l^2$ be defined by

$$f(x) = \begin{cases} |x|e_n & \text{if } 2^{-n-1} < |x| \leq 2^{-n},\ n = 1, 2, \ldots, \\ |x|e_1 & \text{if } |x| > 2^{-1} \text{ or } x = 0, \end{cases}$$

with $e_n = (0, \ldots, 0, 1, 0, \ldots) \in l^2$ (where 1 stands on the nth place). Let $D \subset l^2$ be the cone of all sequences with nonnegative terms $x_0 = 0$, $y_0 = f(x_0) = 0$, and $S = \mathbb{R}$. Then one can prove that $D_c(f + D)(x_0, y_0)(v) = \emptyset\ \forall v \neq 0$. So, by Theorem 5.5, it follows that $x_0 \in \mathrm{Strl}_D(f, S; 1)$.

Example 5.19. (a) Theorem 5.5 is applicable even when the cone D is not pointed (in this sense is better than [10, Theorem 3]). Let us consider the following data: $f : \mathbb{R} \to \mathbb{R}^2$ is given by $f(x) = (|x|, 1/x)$ for $x \neq 0$ and $f(0) = (0, 0)$, $S = \mathbb{R}$, $x_0 = 0$, $y_0 = (0, 0)$, and $D = \{(y_1, y_2) :\ y_1 \geq 0\}$. An easy calculation shows that $D_c(f + D)(x_0, y_0)(v) = (|v|, 0) + D\ \forall\, v \in \mathbb{R}$. Hence, Assumption (B) is satisfied ($(B2)$ is always true for a function), and so $x_0 \in \mathrm{Strl}_D(f, S; 1)$. This conclusion can also be obtained from Proposition 3.10 by choosing $\psi(y_1, y_2) = y_1$. Proposition 3.10 asserts that $x_0 \in \mathrm{Str}_D(f, S; 1)$, i.e., a global minimizer of order one, because x_0 is a global minimizer of order one of $(\psi \circ f)(x) = |x|$. So, it is stronger than the conclusion obtained from Theorem 5.5.

(b) Let us observe that if we take (Example 2 in [10]) $f(x) = (x^2, 1/x^2)$ for $x \neq 0$ and $f(0) = (0, 0)$, and S, x_0, y_0, and D as above, then condition $(B1)$ is not satisfied for $F = f + D$, since $F + D = F$ and $D_c(f + D)(x_0, y_0)(v) = \mathbb{R}_+ \times \mathbb{R}\ \forall\, v \in \mathbb{R}$. Since $f + D$ is D-pseudo Lipschitz at (x_0, y_0) (in fact, f is D-Lipschitz at x_0) and (5.13) is not satisfied, we conclude that $(x_0, y_0) \notin \mathrm{Strl}_D(f + D, S; 1)$ because of Theorem 5.16.

Moreover, using the linear functional $\psi(y_1, y_2) = y_1$ and considering now $F = f$, we see that $x_0 \notin \mathrm{Strl}(\psi \circ f, S; 1)$, and from Proposition 3.12 it follows that $x_0 \notin \mathrm{Strl}_D(f, S; 1)$. But by applying Proposition 3.10, as $x_0 \in \mathrm{Str}(\psi \circ f, S; 2)$, we deduce that $x_0 \in \mathrm{Str}_D(f, S; 2)$. In addition, as f is D-pseudo Lipschitz at (x_0, y_0), (5.13) is satisfied since $D_c f(x_0, y_0)(v) = \emptyset\ \forall v \in \mathbb{R} \setminus \{0\}$ and $x_0 \notin \mathrm{Strl}_D(f, S; 1)$, we conclude that the converse of Theorem 5.16 is not true. On the other hand, we also conclude that we cannot remove the pointedness of D in Corollary 5.17.

Before establishing a sufficient condition when the initial space is not necessarily finite-dimensional, we introduce one notion taken from [27].

The set-valued map F is said to be upper semidifferentiable at $(x_0, y_0) \in \mathrm{Gr}\, F$ [27] if for any sequence $(x_n, y_n) \in \mathrm{Gr}\, F \setminus \{(x_0, y_0)\}$ converging to (x_0, y_0), there exist a subsequence (x_{n_k}, y_{n_k}) and a sequence $t_k \to 0^+$ such that $(\frac{x_{n_k} - x_0}{t_k}, \frac{y_{n_k} - y_0}{t_k}) \to (v, u) \in X \times Y$, with $(v, u) \neq (0, 0)$.

The following theorem largely improves the second part of Theorem 2.1 in Luc [27] in two features: firstly, Theorem 5.20 considers an arbitrary set $S \subset X$ while

Luc's result is for $S = X$, and secondly, the conclusion of Theorem 5.20 provides more information than that in Luc, which asserts only that $(x_0, y_0) \in \mathrm{LMin}_D(F, X)$.

THEOREM 5.20. *Let* $F : X \to 2^Y$, $x_0 \in S \subset X$, $(x_0, y_0) \in \mathrm{Gr}\, F$, $y_0 \in \mathrm{Str}_D\, F(x_0)$ *and assume that* D *is closed and* F *is compact at* x_0 *and upper-semidifferentiable at* (x_0, y_0). *If*

$$D_cF(x_0, y_0)(v) \cap (-D) = \emptyset \quad \forall v \in T(S, x_0) \setminus \{0\}, \tag{5.16}$$

$$D_cF(x_0, y_0)(0) \cap (-D) = \{0\}, \tag{5.17}$$

then $(x_0, y_0) \in \mathrm{Strl}_D(F, S; 1)$.

Proof. Suppose that $(x_0, y_0) \notin \mathrm{Strl}_D(F, S; 1)$, then by Proposition 3.4, there exist sequences $x_n \in S \setminus \{x_0\}$, $y_n \in F(x_n)$, and $d_n \in D$ such that $x_n \to x_0$ and

$$\lim_{n\to\infty} \frac{y_n - y_0 + d_n}{\|x_n - x_0\|} = 0. \tag{5.18}$$

Since F is compact at x_0, there is a subsequence $(x_{n_k}, y_{n_k}) \to (x_0, \bar{y})$, with $\bar{y} \in F(x_0)$. From (5.18), we deduce that $\lim_{n\to\infty}(y_n - y_0 + d_n) = 0$ because $\|x_n - x_0\| \to 0$. As $\lim_{k\to\infty} y_{n_k} = \bar{y}$, we have $\bar{y} - y_0 = \lim_{k\to\infty}(-d_{n_k}) \in -D$ because D is closed. Since $y_0 \in \mathrm{Str}_D\, F(x_0)$ and $\bar{y} \in F(x_0)$, we deduce that $\bar{y} = y_0$. Using the upper-semidifferentiability of F, there is a sequence $t_k \to 0^+$ such that

$$\left(\frac{x_{n_k} - x_0}{t_k}, \frac{y_{n_k} - y_0}{t_k}\right) \to (v, u), \qquad \text{with } (v, u) \neq (0, 0).$$

From here, $v \in T(S, x_0)$ and

$$\frac{\|x_{n_k} - x_0\|}{t_k} \to \|v\|. \tag{5.19}$$

Using the equality

$$\frac{y_{n_k} - y_0 + d_{n_k}}{t_k} = \frac{y_{n_k} - y_0 + d_{n_k}}{\|x_{n_k} - x_0\|} \cdot \frac{\|x_{n_k} - x_0\|}{t_k},$$

from (5.18) and (5.19) it follows that $\lim_{k\to\infty} t_k^{-1}(y_{n_k} - y_0 + d_{n_k}) = 0$. Therefore, $u = \lim_{k\to\infty} t_k^{-1}(y_{n_k} - y_0) = \lim_{k\to\infty} -t_k^{-1} d_{n_k} \in -D$. As $u \in D_cF(x_0, y_0)(v)$, we have a contradiction to (5.16) if $v \neq 0$. If $v = 0$, from (5.17) it follows that $u = 0$, and this contradicts the condition $(v, u) \neq (0, 0)$. □

COROLLARY 5.21. *If* X *and* Y *are finite-dimensional,* $f : X \to Y$ *is continuous at* $x_0 \in S \subset X$, D *is closed, and* (5.16)–(5.17) *hold for* $F = f$, *then* $x_0 \in \mathrm{Strl}_D(f, S; 1)$.

Proof. As X and Y are finite-dimensional, from Proposition 1.1 in [27] it follows that f is upper-semidifferentiable at $(x_0, f(x_0))$. The continuity of f at x_0 implies that f is compact at $(x_0, f(x_0))$ and now the result follows from Theorem 5.20. □

This corollary can be also deduced from Theorem 5.9 taking into account Remark 5.10(2).

Acknowledgments. The authors are grateful to the anonymous referees for their constructive criticism and helpful comments and suggestions.

REFERENCES

[1] T. Amahroq and L. Thibault, *On proto-differentiability and strict proto-differentiability of multifunctions of feasible points in perturbed optimization problems,* Numer. Funct. Anal. Optim., 16 (1995), pp. 1293–1307.

[2] J. P. Aubin, *Contingent derivatives of set-valued maps and existence of solutions to nonlinear inclusions and differentiable inclusion,* in Advances in Mathematics, Supplementary studies, L. Nachbin, ed., Academic Press, New York, 1981, pp. 160–229.

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

[4] A. Auslender, *Stability in mathematical programming with nondifferentiable data,* SIAM J. Control Optim., 22 (1984), pp. 239–254.

[5] E.M. Bednarczuk, *A note on lower semicontinuity of minimal points,* Nonlinear Anal., 50 (2002), pp. 285–297.

[6] E.M. Bednarczuk, *Weak sharp efficiency and growth condition for vector-valued functions with applications,* Optimization, 53 (2004), pp. 455–474.

[7] E.M. Bednarczuk and W. Song, *Contingent epiderivative and its applications to set-valued optimization,* Control Cybernet., 27 (1998), pp. 375–386.

[8] G. Bigi, *Componentwise versus global approaches to nonsmooth multiobjective optimization,* J. Ind. Manag. Optim., 1 (2005), pp. 21–32.

[9] W.W. Breckner and G. Kassay, *A systematization of convexity concepts for sets and functions,* J. Convex Anal., 4 (1997), pp. 109–127.

[10] G.P. Crespi, I. Ginchev, and M. Rocca, *First order optimality conditions in set-valued optimization,* Math. Methods Oper. Res., 63 (2006), pp. 87–106.

[11] M. Durea, *First and second-order Lagrange claims for set-valued maps,* J. Optim. Theory Appl., 133 (2007), pp. 111–116.

[12] F. Flores-Bazán, *Optimality conditions in non-convex set-valued optimization,* Math. Methods Oper. Res., 53 (2001), pp. 403–417.

[13] C. Gerth and P. Weidner, *Nonconvex separation theorems and some applications in vector optimization,* J. Optim. Theory Appl., 67 (1990), pp. 297–320.

[14] I. Ginchev, A. Guerraggio, and M. Rocca, *Isolated minimizers and proper efficiency for $C^{0,1}$ constrained vector optimization problems,* J. Math. Anal. Appl., 309 (2005), pp. 353–368.

[15] I. Ginchev, A. Guerraggio, and M. Rocca, *Geoffrion type characterization of higher-order properly efficient points in vector optimization,* J. Math. Anal. Appl., 328 (2007), pp. 780–788.

[16] A. Göpfert, H. Riahi, C. Tammer, and C. Zălinescu, *Variational Methods in Partially Ordered Spaces,* CMS Books Math., Springer, New York, 2003.

[17] M.R. Hestenes, *Optimization Theory. The Finite Dimensional Case,* Robert E. Krieger Publishing, New York, 1981.

[18] E. Hernández and L. Rodríguez-Marín, *Nonconvex scalarization in set optimization with set-valued maps,* J. Math. Anal. Appl., 325 (2007), pp. 1–18.

[19] L.R. Huang, *Separate necessary and sufficient conditions for the local minimum of a function,* J. Optim. Theory Appl., 125 (2005), pp. 241–246.

[20] J. Jahn, *Scalarization in vector optimization,* Math. Program., 29 (1984), pp. 203–218.

[21] J. Jahn and R. Rauh, *Contingent epiderivatives and set-valued optimization,* Math. Methods Oper. Res., 46 (1997), pp. 193–211.

[22] B. Jiménez, *Strict efficiency in vector optimization,* J. Math. Anal. Appl., 265 (2002), pp. 264–284.

[23] B. Jiménez, *Strict minimality conditions in nondifferentiable multiobjective programming,* J. Optim. Theory Appl., 116 (2003), pp. 99–116.

[24] B. Jiménez and V. Novo, *First and second order sufficient conditions for strict minimality in nonsmooth vector optimization,* J. Math. Anal. Appl., 284 (2003), pp. 496–510.

[25] B. Jiménez and V. Novo, *First order optimality conditions in vector optimization involving stable functions,* Optimization, 57 (2008), pp. 449–471.

[26] S.J. Li, X.Q. Yang, and G.Y. Chen, *Nonconvex vector optimization of set-valued mappings,* J. Math. Anal. Appl., 283 (2003), pp. 337–350.

[27] D.T. Luc, *Contingent derivatives of set-valued maps and applications to vector optimization,* Math. Program., 50 (1991), pp. 99–111.

[28] D.V. Luu, *Higher-order necessary and sufficient conditions for strict local Pareto minima in terms of Studniarski's derivatives,* Optimization, 57 (2008), pp. 593–605.

[29] J.P. Penot, *Differentiability of relations and differential stability of perturbed optimization problems,* SIAM J. Control Optim., 22 (1984), pp. 529–551.

[30] L. Qi, *On an extended Lagrange claim,* J. Optim. Theory Appl., 108 (2001), pp. 685–688.

[31] R. Reemtsen, *On level sets and an approximation problem for the numerical solution of a free boundary problem,* Computing, 27 (1981), pp. 27–35.

[32] R.T. Rockafellar and R.J. Wets, *Variational Analysis,* Springer, Berlin, 1998.

[33] Y. Sawaragi, H. Nakayama, and T. Tanino, *Theory of Multiobjective Optimization,* Academic Press, Orlando, 1985.

[34] D.S. Shi, *Contingent derivative of the perturbation map in multiobjective optimization,* J. Optim. Theory Appl., 70 (1991), pp. 385–396.

[35] M. Studniarski, *Necessary and sufficient conditions for isolated local minima of nonsmooth functions,* SIAM J. Control Optim., 24 (1986), pp. 1044–1049.

[36] A. Taa, *Necessary and sufficient conditions for multiobjective optimization problems,* Optimization, 36 (1996), pp. 97–104.

[37] A. Taa, *Set-valued derivatives of multifunctions and optimality conditions,* Numer. Funct. Anal. Optim., 19 (1998), pp. 121–140.

[38] D.E. Ward, *Characterizations of strict local minima and necessary condition for weak sharp minima,* J. Optim. Theory Appl., 80 (1994), pp. 551–571.

[39] A. Zaffaroni, *Degrees of efficiency and degrees of minimality,* SIAM J. Control Optim., 42 (2003), pp. 1071–1086.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 2, pp. 909–924

© 2009 Society for Industrial and Applied Mathematics

A UNIFIED SET-VALUED APPROACH TO CONTROL IMMUNOTHERAPY*

KHALID KASSARA†

Abstract. Immunotherapy is set as an asymptotic target control problem under mixed state-control constraints with tumor dynamics given by a general ODE. Then a set-valued approach based on Aubin viability theory is used to design feedback protocols with which density of cancer cells may decrease to zero. Existence of such protocols involves a condition C on initial data; otherwise it is shown that either cancer cells cannot be eliminated or condition C may be achieved at a certain instant, in which case the above protocols can then be used. In order to illustrate the approach two examples are studied.

Key words. feedback control, set-valued analysis, viability theory, immunotherapy protocols

AMS subject classifications. Primary, 93C15; Secondary, 93B52, 54C60, 92C50

DOI. 10.1137/07070591X

1. Introduction and statement of the problem.

Nowadays, there is a growing recognition that mathematical modeling [10, 20, 21] can play a central role in cancer research. It can guide laboratory investigations and give scientists deeper insight into how tumors develop and spread.

Immunotherapy, also referred to as biological therapy, stands for a treatment that stimulates the body's immune system to produce antibodies to fight cancer or lessen the side effects associated with some cancer treatments. Mathematical models of such a process abound in the literature and are submitted to continuous improvements; see, for instance, [5, 7, 14, 15, 17, 18, 19] and the references therein. They give rise to numerous studies which investigate immunotherapy control by using the well-known methods of control theory.

For instance, in [4, 9] the authors state an adequate objective functional and use Hamilton–Jacobi equations to derive a bang-bang optimal protocol. An ODE system of five equations is considered in [5] as a model for tumor-immune interaction and vaccine, then the immunotherapy control problem is set and solved in the Bolza context. In [16] geometric methods of nonlinear control are applied to deal with a mathematical model for antiangiogenic treatments.

Moreover, we refer to [8], which uses spreading control techniques [12] in order to deal with the PDE model of [17] by seeking to expand the zones without tumor cells to the entire tissue.

Alternatively to the approaches cited above, a set-valued method is applied in [13] to show, in the particular case of the model established in [14], that feedback protocol laws can be provided as selections of a parameter set-valued map. Unfortunately, the method works only for the initial data that satisfy a condition, interpreted there as *the cancer is less developed at the beginning of the therapy.*

The aim of this paper is to investigate a general class of immunotherapy ODE models in the framework of the set-valued approach which was developed by [13].

*Received by the editors October 20, 2007; accepted for publication (in revised form) October 30, 2008; published electronically February 25, 2009. This research was supported by Hassan II Academy of Science and Technology, Morocco.
http://www.siam.org/journals/sicon/48-2/70591.html

†Department of Mathematics and Computer Science, University Hassan II of Casablanca, P.O. Box 5366, Maarif, 20100, Casablanca, Morocco (kassarak@member.ams.org).

Furthermore, it addresses the issue of how initial data can condition existence of protocols with which cancer cells may be eliminated.

Most of the ODE models of immunotherapy, encountered in the literature [5, 7, 13, 14, 19], can be expressed as the nonlinear control system

$$\dot{x} = f(x,\tau) + G(x,\tau)u, \tag{1.1a}$$

$$\dot{\tau} = \tau\psi(x,\tau), \tag{1.1b}$$

where $x \doteq (x_1, \ldots, x_n)', u = (u_1, \ldots, u_p)'$, and $f \doteq (f_1, \ldots, f_n)'$ for integers n, p with $p \leq n$. The x_i's denote densities of cell populations which compete with tumor cells. Among them, an external source of p populations are infused in the cancerous tissue with rates u_i. Density of tumor cells is denoted by τ.

The operator G maps $\mathbb{R}^n_+ \times \mathbb{R}_+$ into $\mathcal{L}(\mathbb{R}^p, \mathbb{R}^n)$. Both functions f_i and ψ map $\mathbb{R}^n_+ \times \mathbb{R}_+$ into $\mathbb{R}$ and are continuous. The initial data are given by

$$\begin{aligned} &x_i(0) = x_i^0 \text{ for each } i = 1, \ldots, n, \\ &\tau(0) = \tau_0, \end{aligned} \tag{1.2}$$

where all numbers x_i^0 and τ_0 are positive.

In this context, a successful immunotherapy consists of finding a protocol u_θ which satisfies

$$u_\theta : [0, \infty) \to K, \tag{1.3a}$$

$$\tau_\theta \text{ decreases on } [0, \infty), \tag{1.3b}$$

$$\lim_{t\to\infty} \tau_\theta(t) = 0, \tag{1.3c}$$

where (x_θ, τ_θ) denotes a solution to system (1.1) for control u_θ and

$$K \doteq \prod_{i=1}^{p} \ [0, u_i^{\max}]$$

for positive numbers $u_i^{\max}$.

Condition (1.3a) allows one to keep the toxicity to the normal tissue acceptable and (1.3c) expresses that cancer cells are destroyed at a terminal time, while (1.3b) may be optional and aims at reducing undesirable effects on the patient.

Throughout this paper, the Euclidean norm is denoted $\|\|$, and $\langle , \rangle$ is the usual inner product. For a vector z we denote by z_i its ith component. Let T be a linear operator, and we denote its adjoint operator by $T^\star$ and its norm by $\|T\|$. Furthermore we consider the notation

$$\nabla_x \psi \doteq \left(\frac{\partial \psi}{\partial x_1}, \ldots, \frac{\partial \psi}{\partial x_n} \right)',$$

such that for each (x, τ) and (y, ξ) in $\mathbb{R}^n_+ \times \mathbb{R}_+$ we have

$$\dot{\psi}(x,\tau)(y,\xi) = \langle \nabla_x \psi(x,\tau), y \rangle + \xi \frac{\partial \psi}{\partial \tau}(x,\tau).$$

Given a set-valued map $Q : D \to 2^{\mathbb{R}^m}$ and an integer m, the mapping $s : D \to \mathbb{R}^m$ is called a selection of Q if $s(p) \in Q(p)$ for each p. The minimal selection of the map Q is given by

$$s^\star(p) = \pi_{Q(p)}(0) \text{ for all } p \in D.$$

Here π stands for the operator of best approximation.

The layout of this paper is as follows: In the next section we give some background on viability theory along with some facts from set-valued analysis, and an exposition of the main elements of our approach is provided in section 3. In section 4 we prove some topological properties of the feedback map. In section 5 we focus on protocol laws, and section 6 contains the main results on immunotherapy control. Finally in section 7 two models of immunotherapy are studied in order to illustrate the results established.

2. Definitions and preliminary results. Let D denote a subset of the Euclidean space $\mathbb{R}^k$ with $k \geq 1$. The contingent cone [6] at $p \in D$ is defined by

$$T_D(p) \doteq \left\{ q \in \mathbb{R}^k \mid \liminf_{h \downarrow 0} \frac{d(p+hq, D)}{h} = 0 \right\},$$

where

$$d(r, D) \doteq \inf_{q \in D} \|q - r\| \quad \text{ for each } r \in \mathbb{R}^k.$$

We need to recall the following facts we will use throughout the paper.

1. If a subset D is closed and convex, then it is a sleek subset. The latter property consists of that the map $T_D : D \to 2^{\mathbb{R}^k}$ is lower semicontinuous (in short, *lsc*), i.e., for each $p \in D$ and any sequence $(p_n)_n \subset D$ which converges to p, then for each $q \in T_D(p)$ there exists a sequence $q_n \in T_D(p_n)$ that converges to q.

2. The result below is due to [2, section 11.2.5].

LEMMA 2.1. *Let $L \subset \mathbb{R}^l$ and $M \subset \mathbb{R}^m$ (for integers l, m) be two closed sleek subsets and let $\varphi : \mathbb{R}^l \to \mathbb{R}^m$ be a continuously differentiable mapping. If $p \in L \cap \varphi^{-1}(M)$ satisfies the transversality condition*

$$\dot\varphi(p) T_L(p) - T_M(\varphi(p)) = \mathbb{R}^m, \tag{2.1}$$

then

$$T_{L \cap \varphi^{-1}(M)}(p) = T_L(p) \cap \dot\varphi(p)^{-1} T_M(\varphi(p)). \tag{2.2}$$

3. Let $\varphi : \mathbb{R}^k \to \mathbb{R}^k$. Then D is said to be locally viable under the system

$$\begin{aligned} \dot\xi &= \varphi(\xi), \\ \xi(0) &= \xi_0 \end{aligned} \tag{2.3}$$

if for all $\xi_0 \in D$ there exist $\bar t > 0$ and a solution to system (2.3), $\bar\xi(\cdot)$ on $[0, \bar t)$ which is viable in D (i.e., satisfies $\bar\xi(t) \in D$ for all t). Such a property can be characterized in terms of contingent subsets.

LEMMA 2.2 (see [1]). *Assume that φ is continuous on the closed subset D. Then D is locally viable under system* (2.3) *if and only if*

$$\varphi(\xi) \in T_D(\xi) \textit{ for each } \xi \in D. \tag{2.4}$$

4. Next, given the control system

$$\dot{\xi} = \varphi(\xi) + B(\xi)w, \\ w \in W(\xi), \tag{2.5}$$

where w takes values in $\mathbb{R}^m$ and denotes the control, $B : \mathbb{R}^k \to \mathcal{L}(\mathbb{R}^m, \mathbb{R}^k)$, and $W(\cdot)$ stands for the set-valued map of constraints, define the feedback map

$$\mathcal{F}(p) \doteq \{w \in W(p) \mid \varphi(p) + B(p)w \in T_D(p)\}. \tag{2.6}$$

Assume that φ and $B(\cdot)$ are continuous on D; then any continuous selection of the feedback map $\mathcal{F}$ provides a control law that leads to a viable solution to system (2.5) in D. This is so for the minimal selection whenever it is continuous. Otherwise we can use [2, Theorem 4.3.2] as follows.

LEMMA 2.3. *Assume that the feedback map $\mathcal{F}(\cdot)$ is* lsc *with nonempty closed convex values. Then system* (2.5) *with feedback control $w = \pi_{\mathcal{F}(\xi)}(0)$ has a locally viable solution in D.*

5. Let F be a set-valued map from $\mathbb{R}^k$ to $\mathbb{R}^k$. Then consider the differential inclusion

$$\left| \begin{array}{l} \dot{z} \in F(z), \\ z(0) = z_0. \end{array} \right. \tag{2.7}$$

DEFINITION 2.4 (see [2]). *The* capture basin *of D under F is denoted by* $\mathrm{capt}_F(D)$ *and stands for the set of all initial states $z_0 \in \mathbb{R}^k$ such that subset D is reached by one solution to differential inclusion* (2.7).

3. A set-valued approach. In a ready manner we verify that condition (1.3b) can be written as $\psi(x(t), \tau(t)) \leq 0$ for each $t \in [0, \infty)$. Thereby it reduces to the viability of the subset

$$D_0 \doteq \left\{(x, \tau) \in \mathbb{R}^n_+ \times \mathbb{R}_+ \mid \psi(x, \tau) \leq 0\right\}.$$

Being inspired by the theory of Lyapunov functions as studied in [1, section 9.2], we introduce the family of subsets

$$D_\nu \doteq \left\{(x, \tau) \in \mathbb{R}^n_+ \times \mathbb{R}_+ \mid \psi(x, \tau) \leq -\nu\right\} \quad \text{for each } \nu \geq 0. \tag{3.1}$$

Now, let $\nu > 0$ and suppose that protocol $u_\theta : [0, \infty) \to K$ leads to a solution (x_θ, τ_θ) which is globally viable in D_ν; then u_θ solves Problem (1.3). Indeed, it is obvious that u_θ and τ_θ satisfy (1.3a) and (1.3b), respectively. As for (1.3c), thanks to (1.1b) we have

$$\tau_\theta(t) = \tau_\theta(0) \exp\left(\int_0^t \psi(x_\theta(s), \tau_\theta(s)) ds\right) \text{ for each } t \geq 0,$$

which yields the estimate

$$0 \leq \tau_\theta(t) \leq \tau_0 \exp(-\nu t) \text{ for each } t \geq 0. \tag{3.2}$$

Remark 3.1. We notice that the parameter ν in formula (3.2) can be interpreted as the average speed of the therapy. The greater it is, the smaller the therapy horizon.

Definition 3.2. *The mapping $\sigma : D_\nu \to K$ is said to be an immunotherapy protocol law (in short,* itp *law) if the feedback control $u = \sigma(x,\tau)$ is a solution to problem* (1.3) *for all $(x_0,\tau_0) \in D_\nu$.*

According to (2.6) the feedback map can be given for each $\nu > 0$ by

$$\mathcal{F}_\nu(x,\tau) \doteq \{u \in K \mid (f(x,\tau) + G(x,\tau)u, \tau\psi(x,\tau))' \in T_{D_\nu}(x,\tau)\} \tag{3.3}$$

for each $(x,\tau) \in D_\nu$.

Thus, thanks to Lemma 2.2, *itp* laws can be provided by selections σ of the map $\mathcal{F}_\nu(\cdot)$ for which system (1.1) with $u = \sigma(x,\tau)$ has a globally viable solution in D_ν.

Next we proceed to express the contingent cone $T_{D_\nu}(\cdot)$. First of all, by considering (3.1) we get $D_\nu = L \cap \psi^{-1}(M_\nu)$, where

$$L \doteq \mathbb{R}^n_+ \times \mathbb{R}_+ \quad \text{and} \quad M_\nu \doteq (-\infty, -\nu].$$

Since L and M_ν are closed and convex, they are sleek subsets as required by Lemma 2.1, and their contingent cones are given by

$$(y,\xi) \in T_L(x,\tau) \iff \left|\begin{array}{l} y_i \geq 0 \text{ if } x_i = 0 \text{ for } i = 1,\dots,n, \\ \xi \geq 0 \text{ if } \tau = 0 \end{array}\right.$$

and

$$z \in T_{M_\nu}(m) \iff z \leq 0 \text{ if } m = -\nu$$

for all $(x,\tau) \in L$ and $m \in M_\nu$. Moreover, we can easily see that the transversality condition in Lemma 2.1 is satisfied whenever the following conditions hold:

$$\psi \text{ is of class } \mathcal{C}^1 \text{ on } D_\nu, \tag{3.4a}$$

and

$$\left|\begin{array}{l} \text{for all } (x,\tau) \in D_\nu \text{ there exists} \\ j \in \{1,\dots,n+1\} \text{ such that } \dfrac{\partial\psi}{\partial\zeta_j}(x,\tau) < 0, \\ \text{where } \zeta = (x,\tau). \end{array}\right. \tag{3.4b}$$

Lemma 3.3. *Let $\nu > 0$ and suppose that condition* (3.4) *is satisfied. Then for each $(x,\tau) \in D_\nu$ we have*

$$(y,\xi) \in T_{D_\nu}(x,\tau) \iff \left|\begin{array}{l} y_i \geq 0 \textit{ if } x_i = 0 \textit{ for } i = 1,\dots,n, \\ \xi \geq 0 \textit{ if } \tau = 0 \textit{ and} \\ \langle \nabla_x\psi(x,\tau), y\rangle + \xi\dfrac{\partial\psi}{\partial\tau}(x,\tau) \leq 0 \textit{ if } \psi(x,\tau) = -\nu. \end{array}\right. \tag{3.5}$$

4. The feedback map. This section involves the main results we need in order to carry out our approach. First of all, we seek to determine a useful expression of the feedback map $\mathcal{F}_\nu(\cdot)$ given by (3.3). Define the functions

$$h(x,\tau) \doteq -G^\star(x,\tau)\nabla_x\psi(x,\tau), \tag{4.1a}$$

$$\ell(x,\tau) \doteq \langle \nabla_x\psi(x,\tau), f(x,\tau)\rangle + \tau\psi(x,\tau)\frac{\partial\psi}{\partial\tau}(x,\tau) \tag{4.1b}$$

and the map

$$C(x,\tau) \doteq \{u \in K \mid \langle h(x,\tau), u\rangle \geq \ell(x,\tau)\} \tag{4.1c}$$

for each $(x,\tau) \in \mathbb{R}^n_+ \times \mathbb{R}_+$. In addition we consider the assumption

$$\left| \begin{array}{l} x = (x_1, \dots, x_n)' \\ x_i = 0 \text{ and } \tau \geq 0 \end{array} \right. \Rightarrow f_i(x,\tau) + (G(x,\tau)u)_i \geq 0 \text{ for all } u \in K \tag{4.2}$$

for all $i = 1, \dots, n$.

Proposition 4.1. *Let $\nu > 0$ be such that* (3.4) *and* (4.2) *hold true. Then for each $(x,\tau) \in D_\nu$ we have*

$$\mathcal{F}_\nu(x,\tau) = \left| \begin{array}{lll} K & \textit{if} & \psi(x,\tau) < -\nu, \\ C(x,\tau) & \textit{if} & \psi(x,\tau) = -\nu. \end{array} \right. \tag{4.3}$$

Proof. First, by (4.1) we remark that

$$\dot{\psi}(x,\tau)(f(x,\tau) + G(x,\tau)u, \tau\psi(x,\tau))' = -\langle h(x,\tau), u\rangle + \ell(x,\tau).$$

By (3.3) and Lemma 3.3 we get

$$u \in \mathcal{F}_\nu(x,\tau) \iff \left| \begin{array}{lll} f_i(x,\tau) + (G(x,\tau)u)_i \geq 0 & \text{if} & x_i = 0 \text{ for } i = 1, \dots, n, \\ \tau\psi(x,\tau) \geq 0 & \text{if} & \tau = 0, \\ -\langle h(x,\tau), u\rangle + \ell(x,\tau) \leq 0 & \text{if} & \psi(x,\tau) = -\nu. \end{array} \right.$$

Thanks to (4.2), it follows that

$$u \in \mathcal{F}_\nu(x,\tau) \iff \langle h(x,\tau), u\rangle \geq \ell(x,\tau) \text{ if } \psi(x,\tau) = -\nu,$$

ending the proof of the proposition. □

Subsequently, let us consider the assumption

$$\left| \begin{array}{l} \text{for each } (x,\tau) \in D_\nu \text{ there exists} \\ u \in K \text{ such that } \langle h(x,\tau), u\rangle > \ell(x,\tau), \end{array} \right. \tag{4.4}$$

where the functions h and ℓ are given by (4.1).

Lemma 4.2. *The map $C(\cdot)$ given by* (4.1c) *is lsc on D_ν whenever condition* (4.4) *is satisfied.*

Proof. We rewrite the map C in the context of [3, Proposition 1.5.2]. We get

$$C(x,\tau) = \{u \in \bar{F}(x,\tau) \mid \bar{f}(x,\tau,u) \in \bar{G}(x,\tau)\},$$

where, for each $(x,\tau) \in D_\nu$, we have

$$\bar{F}(x,\tau) \doteq K, \quad \bar{f}(x,\tau,u) \doteq \langle h(x,\tau), u\rangle, \quad \text{and} \quad \bar{G}(x,\tau) \doteq [\ell(x,\tau), \infty).$$

Thereafter, we easily check the hypotheses of the cited proposition as follows:

(i) The map $\bar{F}$ is lsc with convex values.

(ii) $\bar{f}$ is continuous.

(iii) For all $(x,\tau) \in D_\nu$, the mapping $u \to \bar{f}(x,\tau,u)$ is affine.

(iv) For all (x,τ), $\bar{G}(x,\tau)$ is convex and its interior is nonempty.

(v) The graph of the map $(x,\tau) \in D_\nu \to \operatorname{int}(\bar{G}(x,\tau))$ is open.

(vi) For all $(x,\tau) \in D_\nu$, there exists $u \in \bar{F}(x,\tau)$ such that $\bar{f}(x,\tau,u) \in \operatorname{int}(\bar{G}(x,\tau))$.

Note that (vi) is due to condition (4.4), whence the map C is lsc. □

LEMMA 4.3. *Let $\nu > 0$ be such that condition* (4.4) *holds true. Then the minimal selection of the map C is continuous on D_ν.*

Proof. By Lemma 4.2 the map C is lsc. Then we can use [11, Theorem 4.1] by verifying that the subset

$$K_\epsilon \doteq \{(x,\tau) \in D_\nu \mid \exists u \in C(x,\tau) \text{ s.t. } \|u\| \leq \epsilon\}$$

is closed in D_ν for all $\epsilon > 0$. Indeed, let $((x_n,\tau_n))_n$ be a sequence in K_ϵ which converges to $(\bar{x},\bar{\tau})$. Then there exists a sequence $(u_n)_n \subset K$ such that

$$\|u_n\| \leq \epsilon \quad \text{and} \quad \langle h(x_n,\tau_n), u_n\rangle \geq \ell(x_n,\tau_n) \text{ for all } n. \tag{4.5}$$

Now, as $(u_n)_n$ is bounded it has a subsequence $(u_k)_k$ which converges to $\bar{u}$. Then by noting that h and ℓ are continuous and letting $k \to \infty$ in (4.5), we get

$$\|\bar{u}\| \leq \epsilon \quad \text{and} \quad \langle h(\bar{x},\bar{\tau}), \bar{u}\rangle \geq \ell(\bar{x},\bar{\tau}).$$

This implies that $(\bar{x},\bar{\tau}) \in K_\epsilon$ and therefore K_ϵ is closed. □

LEMMA 4.4. *Let $\nu > 0$ and suppose that condition* (4.4) *holds true. Then the map $\mathcal{F}_\nu(\cdot)$ as given by* (4.3) *is lsc on D_ν.*

Proof. Let $(x_n,\tau_n)_n$ be a sequence of D_ν that converges to $(x,\tau) \in D_\nu$ and $u \in \mathcal{F}_\nu(x,\tau)$. We have to seek a sequence $(u_n)_n$ that satisfies

$$\left| \begin{array}{l} u_n \in \mathcal{F}_\nu(x_n,\tau_n) \text{ for each } n, \\ \text{and } u_n \to u. \end{array} \right. \tag{4.6}$$

Assume that $\psi(x,\tau) < -\nu$. Since the function ψ is continuous and $(x_n,\tau_n) \to (x,\tau)$ we can consider the smallest number n_0 such that

$$\psi(x_n,\tau_n) < -\nu \text{ for all } n \geq n_0.$$

Then the sequence defined by

$$u_n \doteq \left| \begin{array}{ll} u & \text{if } n \geq n_0, \\ v_n & \text{if } n < n_0, \end{array} \right.$$

where

$$v_n \in C(x_n,\tau_n) \text{ for all } n < n_0,$$

merely satisfies (4.6) due to the fact that $\psi(x_n,\tau_n) = -\nu$ whenever $n < n_0$.

Now suppose that $\psi(x,\tau) = -\nu$; then $u \in C(x,\tau)$. By Lemma 4.2 the map $C(\cdot)$ is lsc as condition (4.4) holds true. It follows that there exists a sequence $(u_n)_n$ such that $u_n \in C(x_n,\tau_n)$ for each n and $u_n \to u$. Thanks to (4.3) we get $u_n \in \mathcal{F}_\nu(x_n,\tau_n)$ for all n, as required in (4.6). □

5. Immunotherapy protocol laws. This section is entirely dedicated to itp laws. We know from section 3 that these laws can be provided as selections of the feedback map $\mathcal{F}_\nu(\cdot)$, which lead to global solutions to the system. For that end, we consider the linear growth assumption on D_ν,

$$\|f(x,\tau)\| \leq m_1(\tau)(\|x\|+1) \quad \text{and} \quad \|G(x,\tau)\| \leq m_2(\tau), \tag{5.1}$$

where m_1 and m_2 denote positive functions that map bounded subsets into bounded images.

LEMMA 5.1. *Assume that* (5.1) *holds and let* $\sigma : D_\nu \to K$ *be such that feedback control* $u = \sigma(x,\tau)$ *leads to a locally viable solution* $(\bar{x},\bar{\tau})$ *to system* (1.1) *for all* $(x_0,\tau_0) \in D_\nu$. *Then that solution is global for all* $(x_0,\tau_0) \in D_\nu$.

Proof. Let $(\bar{x},\bar{\tau})$ be defined over a maximal interval $[0,t_1)$. We have to show that $t_1 = \infty$. Indeed, assume that $t_1 < \infty$. As the nonnegative function $\bar{\tau}$ is decreasing on $[0,t_1)$ it follows that $\bar{\tau}(t) \leq \tau_0$ for all $t \in [0,t_1)$. Then there exists $\bar{m} > 0$ such that

$$m_1(\tau(t)) \leq \bar{m} \quad \text{and} \quad m_2(\tau(t)) \leq \bar{m} \text{ on } [0,\bar{t})\,.$$

It follows that the right-hand side of (1.1a) satisfies the estimate

$$\|f(x,\tau) + G(x,\tau)\sigma(x,\tau)\| \leq \bar{m}\left(\|x\| + \sup_{0\leq i\leq p}(u_i^{\max}) + 1\right),$$

which yields a linear growth for (1.1a). This implies that

$$\bar{x}(t) \to x_1 \text{ when } t \to t_1.$$

As $\bar{\tau}(\cdot)$ is a nonnegative decreasing function, we have

$$\bar{\tau}(t) \to \tau_1 \text{ when } t \to t_1.$$

Therefore

$$(\bar{x}(t),\bar{\tau}(t)) \to (x_1,\tau_1) \text{ when } t \to t_1,$$

and $(x_1,\tau_1) \in D_\nu$ because D_ν is closed. Now, by considering (x_1,τ_1) as an initial data, it follows that system (1.1) admits a viable solution in D_ν, which starts from (x_1,τ_1) at time t_1, contradicting the fact that the interval $[0,t_1)$ is maximal. □

For all $\nu > 0$, the minimal selection of the map $\mathcal{F}_\nu(\cdot)$ of (4.3) is given for all $(x,\tau) \in D_\nu$ by

$$s_\nu^\star(x,\tau) \doteq \left|\begin{array}{lll} 0 & \text{if} & \psi(x,\tau) < -\nu, \\ \pi_{C(x,\tau)}(0) & \text{if} & \psi(x,\tau) = -\nu. \end{array}\right. \tag{5.2}$$

Although $s_\nu^\star$ is not continuous, we will see next that it can provide an itp law.

THEOREM 5.2. *Let* $\nu > 0$ *and assume both conditions* (4.4) *and* (5.1). *Then* $s_\nu^\star$ *given by* (5.2) *stands for an itp law.*

Proof. Lemma 4.4 implies that $\mathcal{F}_\nu(\cdot)$ is lsc. Since it has closed convex values, we can use Lemma 2.3. Thus, by using feedback control $u = s_\nu^\star(x,\tau)$, system (1.1) has a local solution which is viable in D_ν. Now, by virtue of Lemma 5.1, $s_\nu^\star(\cdot)$ is an itp law. □

We now establish a result on continuous itp laws.

THEOREM 5.3. *Assume that conditions* (4.4) *and* (5.1) *are satisfied. Let* σ *be a continuous selection of the map* C *and let* $\zeta : \mathbb{R}^+ \to [0,1]$ *be continuous such that* $\zeta(0) = 1$. *Then a continuous itp law can be given by*

$$s_\nu(x,\tau) \doteq \zeta(-\psi(x,\tau) - \nu)\sigma(x,\tau) \quad \textit{for each } (x,\tau) \in D_\nu. \tag{5.3}$$

Proof. We notice first that s_ν is continuous as being the composite of continuous functions. Since $0 \le \zeta(-\psi(x,\tau) - \nu) \le 1$ for all $(x,\tau) \in D_\nu$, then $s_\nu(x,\tau) \in K$ if $\psi(x,\tau) < -\nu$. Otherwise we get $s_\nu(x,\tau) = \sigma(x,\tau) \in C(x,\tau)$, whence s_ν is a selection of the map $\mathcal{F}_\nu(\cdot)$. Now we use Lemma 5.1 to conclude that s_ν is an itp law. □

Remark 5.4. As an application of Theorem 5.3, let $\zeta(z) = e^{-\mu z}$ for some positive parameter μ and let σ stand for the minimal selection of the map C (continuous thanks to Lemma 4.3). Then we get an important family of continuous itp laws on D_ν as follows:

$$\sigma_\nu(x,\tau) \doteq e^{\mu(\psi(x,\tau)+\nu)}\pi_{C(x,\tau)}(0) \quad \text{for each } (x,\tau) \in D_\nu. \tag{5.4}$$

We emphasize that these laws are slightly higher than the minimal itp law $s^\star_\nu$ given by (5.2). This is due to the exponential decay in formula (5.4).

6. Immunotherapy control. In this section we examine the implications of our set-valued approach on the treatment of the immunotherapy control problem (1.3a)–(1.3c). First of all, we define the set-valued map

$$F(x,\tau) \doteq \{(f(x,\tau) + G(x,\tau)u, \tau\psi(x,\tau))' \mid u \in K\} \tag{6.1a}$$

for all $(x,\tau) \in \mathbb{R}^n_+ \times \mathbb{R}_+$. Therefore solutions to system (1.1) are also given as solutions to the differential inclusion

$$\left| \begin{array}{l} (\dot{x},\dot{\tau}) \in F(x,\tau), \\ x(0) = x_0, \tau(0) = \tau_0, \end{array} \right. \tag{6.1b}$$

and vice versa. Then we let

$$\Sigma \doteq \Omega_+ \cap \mathrm{capt}_F(\Omega_-), \tag{6.2}$$

where the subset $\mathrm{capt}_F(\cdot)$ stands for the capture basin as provided by Definition 2.4, and the subsets Ω_- and Ω_+ are given by

$$\Omega_- \doteq \left\{(x,\tau) \in \mathbb{R}^n_+ \times \mathbb{R}_+ \mid \psi(x,\tau) < 0\right\} \tag{6.3a}$$

and

$$\Omega_+ \doteq \left\{(x,\tau) \in \mathbb{R}^n_+ \times \mathbb{R}_+ \mid \psi(x,\tau) \ge 0\right\}. \tag{6.3b}$$

Then we can state the following result.

THEOREM 6.1. *Consider system* (1.1) *with initial data* (x_0,τ_0), *and let* $\nu_0 \doteq -\psi(x_0,\tau_0)$. *Then either of the following situations may hold:*

(i) $(x_0,\tau_0) \in \Omega_-$*: in this case* $s^\star_{\nu_0}(\cdot)$ *and* $s_{\nu_0}(\cdot)$ *as given by* (5.2) *and* (5.3), *respectively, are itp laws whenever conditions* (4.4) *and* (5.1) *hold true for* ν_0.

(ii) $(x_0,\tau_0) \in \Sigma$*: if condition* (4.4) *is satisfied on* D_0, *then a protocol which satisfies* (1.3a) *and* (1.3c) *exists, and it is given by an itp law after an instant* t_{im}.

(iii) $(x_0, \tau_0) \in \Omega_+ \setminus \Sigma$*: then there is no protocol which solves problem* (1.3a)–(1.3c).

Proof. (i) We are in a position to apply Theorems 5.2 and 5.3, respectively.

(ii) Since $(x_0, \tau_0) \in \operatorname{capt}_F(\Omega_-)$ then there exists a solution $(\bar{x}, \bar{\tau})$ to differential inclusion (6.1b), which satisfies

$$(\bar{x}(t_{\mathrm{im}}), \bar{\tau}(t_{\mathrm{im}})) \in \Omega_- \text{ at an instant } t_{\mathrm{im}}.$$

Let $\bar{u} : [0, \infty) \to K$ be a control which leads to such a solution in accord with (6.1). Let $x_{\mathrm{im}} \doteq \bar{x}(t_{\mathrm{im}})$ and $\tau_{\mathrm{im}} \doteq \bar{\tau}(t_{\mathrm{im}})$. Then as $(x_{\mathrm{im}}, \tau_{\mathrm{im}}) \in \Omega_-$ one can use (i) to get an itp law for $(x_{\mathrm{im}}, \tau_{\mathrm{im}})$ as initial data, say $s_{\mathrm{im}} : D_{\nu_m} \to K$, with $\nu_m \doteq -\psi(x_{\mathrm{im}}, \tau_{\mathrm{im}})$. Subsequently, a control that solves problem (1.3a)–(1.3c) can be given by

$$\left| \begin{array}{lll} \bar{u}(t) & \text{if} & 0 \le t < t_{\mathrm{im}}, \\ s_{\mathrm{im}}(x(t), \tau(t)) & \text{if} & t \ge t_{\mathrm{im}}. \end{array} \right. \tag{6.4}$$

(iii) For such initial data (x_0, τ_0) all solutions $(\bar{x}, \bar{\tau})$ to differential inclusion (6.1b) are viable in subset Ω_+, that is to say,

$$\psi(\bar{x}(t), \bar{\tau}(t)) \ge 0 \text{ for all } t \ge 0. \tag{6.5}$$

As a result, any control u taking values in K will lead to such a solution. Now by using (1.1b) and (6.5) we get

$$\bar{\tau}(t) \ge \tau_0 \text{ for each } t \ge 0;$$

therefore $\bar{\tau}(t) \not\to 0$ when $t \to \infty$. □

Next, we show how protocol $\bar{u}$ and instant t_{im}, which are involved by the proof of Theorem 6.1 (ii), can be determined. For each $\alpha > 0$, let us define the set-valued map

$$C_\alpha(x, \tau) \doteq \{v \in K \mid \langle h(x, \tau), v\rangle - \ell(x, \tau) \ge \alpha\} \tag{6.6}$$

for all $(x, \tau) \in \mathbb{R}^n_+ \times \mathbb{R}_+$, where h and ℓ are as in (4.1a) and (4.1b).

Proposition 6.2. *Let* $(x_0, \tau_0) \in \Omega_+$ *and assume the statements below:*

(i) σ *is a continuous selection of the map* $C_\alpha(\cdot)$ *given by* (6.6) *for some* $\alpha > 0$.

(ii) *System* (1.1) *with feedback control* $u = \sigma(x, \tau)$ *admits a solution on an interval* $[0, t_{\mathrm{im}}]$*, where* t_{im} *satisfies*

$$t_{\mathrm{im}} > \frac{\psi(x_0, \tau_0)}{\alpha}. \tag{6.7}$$

Then the protocol given by $u \doteq \sigma(x, \tau)$ *steers system* (1.1) *from* (x_0, τ_0) *to* Ω_- *at time* t_{im}*, that is,* $(x_0, \tau_0) \in \Sigma$.

Proof. Let $(\bar{x}, \bar{\tau})$ denote the solution which is provided by (ii). Then we get

$$\psi(\bar{x}(t_{\mathrm{im}}), \bar{\tau}(t_{\mathrm{im}})) = \psi(x_0, \tau_0) + \int_0^{t_{\mathrm{im}}} \left[\langle \nabla_x \psi(\bar{x}(s), \bar{\tau}(s)), \dot{\bar{x}}(s)\rangle + \dot{\bar{\tau}}(s) \frac{\partial \psi}{\partial \tau}(\bar{x}(s), \bar{\tau}(s)) \right] ds.$$

Next, by putting $\bar{u} \doteq \sigma(\bar{x}, \bar{\tau})$, we use formulas (4.1a) and (4.1b) to get

$$\psi(\bar{x}(t_{\mathrm{im}}), \bar{\tau}(t_{\mathrm{im}})) = \psi(x_0, \tau_0) - \int_0^{t_{\mathrm{im}}} \left[\langle h(\bar{x}(s), \bar{\tau}(s)), \bar{u}(s)\rangle - \ell(\bar{x}(s), \bar{\tau}(s)) \right] ds.$$

Since σ is a selection of the map $C_\alpha(\cdot)$ then (6.6) yields

$$\psi(\bar{x}(t_{\text{im}}), \bar{\tau}(t_{\text{im}})) \leq \psi(x_0, \tau_0) - \alpha t_{\text{im}}.$$

Thanks to (6.7) it follows that $\psi(\bar{x}(t_{\text{im}}), \bar{\tau}(t_{\text{im}})) < 0$. □

Remark 6.3. We can adapt Lemmas 4.3 and 4.2 to get similar results relative to the map $C_\alpha(\cdot)$. We need to replace condition (4.4) by

$$\left| \begin{array}{l} \text{For all } (x,\tau) \in \mathbb{R}_+^n \times \mathbb{R}_+ \text{ there exists} \\ u \in K \text{ such that } \langle h(x,\tau), u\rangle - \ell(x,\tau) > \alpha, \end{array} \right.$$

and then use $\alpha + \ell$ instead of ℓ.

Ultimately, as a noteworthy fact in cancer modeling, the results above show that any one of the three instances below may arise for a patient having a cancer at the stage (x_0, τ_0) (see Figure 6.1):

(A) $(x_0, \tau_0) \in \Omega_-$: This means that the tumor is less developed with respect to the immune system. It can be cured with an itp law as derived from Theorem 6.1 (i). The notable advantage is that tumor cells will decrease during the therapy, in keeping with the patient's quality of life. Moreover, one may either use the minimal itp law (5.2) in order to reduce the amounts of the administered cells or else be interested in the continuous protocols provided in Remark 5.4 whenever smoothness is required to avoid undesirable effects.

(B) $(x_0, \tau_0) \in \Sigma$: The cancer is more developed. The protocol law σ given by Proposition 6.2 will bring the cancer to a better stage $(x_1, \tau_1) \in \Omega_-$ at an instant t_{im}. This hereby allows one to use an itp law after t_{im}, as in instance (A). Note that a sudden change in the variation of tumor cells may occur within horizon t_{im}, as their density is not necessarily decreasing.

(C) $(x_0, \tau_0) \in \Omega_+ \setminus \Sigma$: The cancer is so advanced that it is not curable, as shown by Theorem 6.1 (iii).

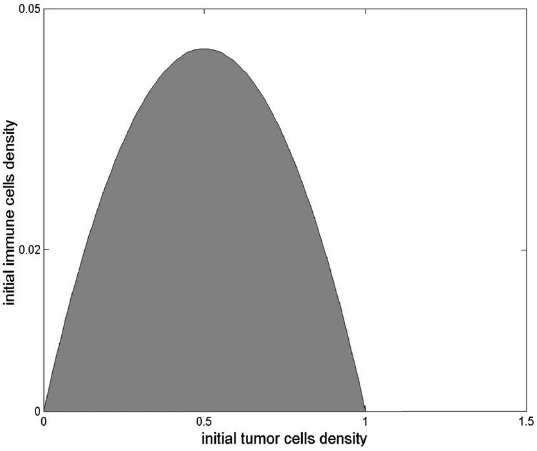


FIG. 6.1. *In gray: plot of the zone* Ω_+ *which corresponds to immunotherapy model* (7.1), *described by both cases* (B) *and* (C) *above. The complimentary zone provides instance* (A).

7. Examples. We begin by investigating the model of [14], which is given by

$$\dot{y} = c\tau - \mu_2 y + \frac{p_1 y z}{g_1 + z} + s_1 u_1, \tag{7.1a}$$

$$\dot{z} = \frac{p_3 y \tau}{g_3 + \tau} - \mu_3 z + s_2 u_2, \tag{7.1b}$$

$$\dot{\tau} = r_2 \tau (1 - b\tau) - \frac{a y \tau}{g_2 + \tau}, \tag{7.1c}$$

with normalized initial conditions

$$y(0) = 1, \; z(0) = 1, \; \tau(0) = 1, \tag{7.1d}$$

where $y(\cdot)$ stands for density of the activated immune system cells, or effector cells (ECs). Concentration of Interleukin-2 (IL-2) is denoted by $z(\cdot)$ and tumor cells (TCs) by $\tau(\cdot)$. All parameters are positive constants; see [14] for their values and units.

In the first differential equation the parameter c models the antigenicity of the tumor, the second term represents the natural death of the effector cells at the rate of μ_2, the third term is of Michaelis–Menton form to indicate the saturated effects of the immune response whereby effector cells are stimulated by IL-2, and the final term involves the strength of the treatment s_1 and the control $u_1(\cdot)$ that represent an external source of ECs.

The second equation gives the rate of change for the concentration of IL-2, the IL-2 source is modeled by another Michaelis–Menton term in which the TCs stimulate the interaction with the ECs to produce more IL-2, the second term represents the loss of these cells at the rate of μ_3, and the last term involves both the strength of the treatment s_2 and the supply rate of IL-2, $u_2(\cdot)$.

The third equation includes a logistic term in order to model the rate of change of TCs. The loss of tumor cells is represented by a Michaelis–Menton term to indicate the limited interaction between the tumor and ECs.

We then see that model (7.1) is involved in our approach by taking $n = p = 2$, and the functions f, ψ, and G in system (1.1) are given by

$$f(x,\tau) \doteq \left(c\tau - \mu_2 y + \frac{p_1 y z}{g_1 + z} \; , \; \frac{p_3 y \tau}{g_3 + \tau} - \mu_3 z \right)'$$

and

$$\psi(x,\tau) \doteq r_2(1 - b\tau) - \frac{a y}{g_2 + \tau}, \qquad G(x,\tau) \doteq \operatorname{diag}(s_1, s_2)$$

for all $(x,\tau) \in \mathbb{R}_+^2 \times \mathbb{R}_+$ and $x = (y,z)'$. We have

$$\nabla_x \psi(x,\tau) = \left(-\frac{a}{g_2 + \tau} \; , \; 0 \right)',$$

$$\frac{\partial \psi}{\partial \tau}(x,\tau) = -r_2 b + \frac{a y}{(g_2 + \tau)^2}$$

for each $(x,\tau) \in \mathbb{R}_+^2 \times \mathbb{R}_+$. It follows that condition (3.4b), which is needed in Lemma 3.3, is satisfied because

$$\frac{\partial \psi}{\partial y}(x,\tau) = -\frac{a}{g_2 + \tau} < 0.$$

The functions h and ℓ of formulas (4.1) are given by

$$h(x,\tau) \doteq \left(\frac{s_1 a}{g_2+\tau} \ , \ 0 \right)'$$

and

$$\ell(x,\tau) \doteq -\frac{a}{g_2+\tau}\left(c\tau - \mu_2 y + \frac{p_1 yz}{g_1+z} \right) \\ + \tau\left(r_2(1-b\tau) - \frac{ay}{g_2+\tau} \right)\left(-r_2 b + \frac{ay}{(g_2+\tau)^2} \right).$$

Then the map C of (4.1c) can be expressed by

$$C(x,\tau) \doteq \left\{ u \in [0, u^1_{\max}] \mid uh(x,\tau) \geq \ell(x,\tau) \right\} \times \left[0, u^2_{\max}\right],$$

and thereby the minimal selection of $\mathcal{F}_\nu(\cdot)$ is as follows:

$$s^\star_\nu(x,\tau) \doteq \left| \begin{array}{lll} (0,0)' & \text{if} & \psi(x,\tau) < -\nu, \\ (\varrho(x,\tau),0)' & \text{if} & \psi(x,\tau) = -\nu, \end{array} \right.$$

where

$$\varrho(x,\tau) \doteq \max\left(-\frac{1}{s_1}\left(c\tau - \mu_2 y + \frac{p_1 yz}{g_1+z} \right) - \frac{\nu\tau}{s_1 a}\left(-r_2 b(g_2+\tau) + \frac{ay}{(g_2+\tau)} \right), 0 \right).$$

As a result, the map $\mathcal{F}_\nu(\cdot)$ admits a continuous selection given by

$$s_\nu(x,\tau) \doteq \left(\min\left(\frac{(g_2+\tau)\exp(\psi(x,\tau)+\nu)\max(\ell(x,\tau),0)}{s_1 a}, u_1^{\max} \right), 0 \right)'.$$

As for condition (5.1), it is also satisfied by taking

$$m_1(\tau) \doteq \max(p_1 + p_3 + \mu_2, \mu_3, c\tau) \quad \text{and} \quad m_2(\tau) \doteq \max(s_1, s_2).$$

It follows that the conditions of Theorem 6.1 hold true whenever $(x_0, \tau_0) \in D_\nu$.

We now turn to examine the family of immunotherapy models studied in [7]:

$$\dot{x} = \beta(\tau)x - \mu(\tau)x + \sigma q(\tau) + u(t), \tag{7.2a}$$

$$\dot{\tau} = \tau(g(\tau) - \phi(\tau)x), \tag{7.2b}$$

where $x(\cdot)$ and $\tau(\cdot)$, respectively, stand for the densities of ECs and TCs, and $g(\tau)$ summarizes many widely used models of tumor growth rates, such as the Stepanova model

$$\left| \begin{array}{l} g(\tau) \doteq \alpha > 0, \ \phi(\tau) \doteq 1, \ \beta(\tau) \doteq \beta_1\tau, \\ q(\tau) \doteq 1 \text{ and } \mu(\tau) \doteq \mu_0 + \mu_2\tau^2. \end{array} \right.$$

As regards the de Vladar–Gonzalez model, it is similar except that

$$g(\tau) \doteq \alpha\log(K/\tau).$$

The Kuznetsor model consists of taking

$$\left|\begin{array}{l} g(\tau) \doteq \alpha(1-\tau/K),\ \phi(\tau) \doteq 1,\ \beta(\tau) \doteq \beta_\infty \tau/(m+\tau), \\ \mu(\tau) \doteq \mu(0) + \mu_1\tau,\ \text{and } q(\tau) \doteq 1, \end{array}\right.$$

We can easily see that system (7.2) represents a particular case of (1.1) where the functions f, ψ, and G are given as

$$f(x,\tau) \doteq \beta(\tau)x - \mu(\tau)x + \sigma q(\tau), \quad G(x,\tau) \doteq 1,$$

and

$$\psi(x,\tau) \doteq g(\tau) - \phi(\tau)x$$

for all $(x,\tau) \in \mathbb{R}_+ \times \mathbb{R}_+$. Thereby we can apply our set-valued approach by proceeding as in the preceding example. The partial derivatives of ψ are given by

$$\frac{\partial\psi}{\partial x}(x,\tau) = -\phi(\tau)$$

and

$$\frac{\partial\psi}{\partial\tau}(x,\tau) = \dot g(\tau) - \dot\phi(\tau)x.$$

We can see that conditions (4.2) and (3.4b) are satisfied. Consequently we get

$$\ell(x,\tau) \doteq -\phi(\tau)(\beta(\tau)x - \mu(\tau)x + \sigma q(\tau)) + \tau(\dot g(\tau) - \dot\phi(\tau)x)(g(\tau) - \phi(\tau)x)$$

and

$$h(x,\tau) \doteq \phi(\tau).$$

Then the map C of (4.1c) is given by

$$C(x,\tau) \doteq \{u \in [0,u_{\max}] \mid u\phi(\tau) \geq \ell(x,\tau)\}.$$

Therefore the minimal *itp* law can be expressed by

$$s^\star_\nu(x,\tau) \doteq \left|\begin{array}{lll} 0 & \text{if} & \psi(x,\tau) < -\nu, \\ \varrho(x,\tau) & \text{if} & \psi(x,\tau) = -\nu, \end{array}\right.$$

where

$$\varrho(x,\tau) \doteq \min\left(\max\left(\frac{\ell(x,\tau)}{\phi(\tau)},0\right),u_{\max}\right).$$

We can use Theorem 5.3 to get the continuous itp law

$$s_\nu(x,\tau) \doteq \exp(\psi(x,\tau)+\nu)\varrho(x,\tau). \tag{7.3}$$

In addition, we notice that condition (5.1) is also fulfilled. In Figures 7.1, 7.2, and 7.3 we summarize numerical results of the Kuznetsov model that represents a particular case of system (7.2) in which

$$\left|\begin{array}{l} g(\tau) = 1.636(1-\tau/100),\ \beta(\tau) = 1.131\tau/(20.19+\tau), \\ \mu(\tau) = 0.347 + 0.0311\tau,\ q(\tau) = 1,\ \text{and } \sigma = 0.6, \end{array}\right.$$

with the initial cells densities given by

$$x_0 = 1 \quad \text{and} \quad \tau_0 = 70,$$

where the function ψ takes the negative value -0.5092. We use the law $s_{0.5092}$ given by (7.3).

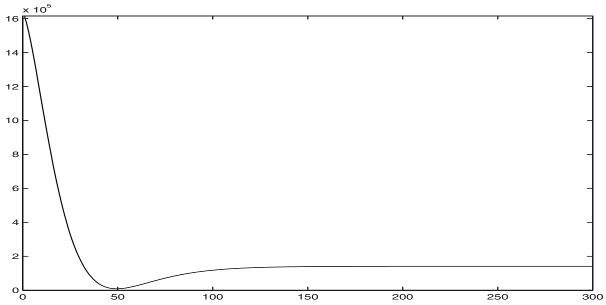


FIG. 7.1. *Dose of infused ECs.*

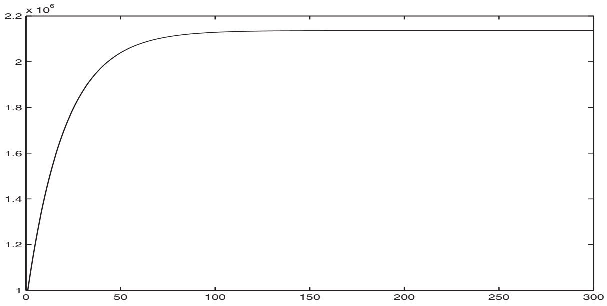


FIG. 7.2. *Density of ECs.*

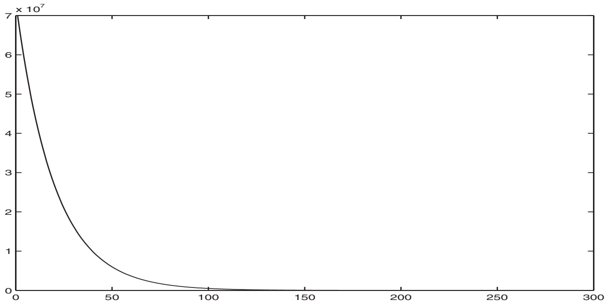


FIG. 7.3. *Density of TCs.*

Acknowledgment. The author is grateful to unknown reviewers for their constructive criticism and helpful suggestions.

REFERENCES

[1] J.-P. AUBIN, *Viability Theory*, Birkhäuser, Boston, 1991.
[2] J.-P. AUBIN, *Dynamic Economic Theory: A Viability Approach*, Springer, Berlin, 1997.
[3] J.-P. AUBIN AND H. FRANKOWSKA, *Set-Valued Analysis*, Birkhäuser, Boston, 1990.
[4] T. BURDEN, J. ERNSTBERGER, AND K. R. FISTER, *Optimal control applied to immunotherapy*, Discrete Contin. Dyn. Syst. Ser. B, 4 (2004), pp. 135–146.
[5] F. CASTIGLIONE AND B. PICCOLI, *Cancer immunotherapy, mathematical modeling and optimal control*, J. Theoret. Biol., 247 (2007), pp. 723–732.
[6] V. DEIMLING, *Multivalued Differential Equations*, Walter de Gruyter, Berlin, 1992.

[7] A. D'Onofrio, *Metamodeling tumor-immune system interaction, tumor evasion and immunotherapy*, Math. Comput. Modelling, 47 (2008), pp. 614–637.

[8] A. El Jai and K. Kassara, *Target control by using feedback spreading control with application to immunotherapy*, Internat. J. Control, 79 (2006), pp. 813–821.

[9] K. R. Fister and H. Donnelly, *Immunotherapy: An optimal control theory approach*, Math. Biosci. Eng., 2 (2005), pp. 499–510.

[10] A. Friedman, *A hierarchy of cancer models and their mathematical challenges*, Discrete Contin. Dyn. Syst. Ser. B, 4 (2004), pp. 147–159.

[11] V. Gutev and S. Nedev, *Continuous selections and reflexive Banach spaces*, Proc. Amer. Math. Soc., 129 (2001), pp. 1853–1860.

[12] K. Kassara, *Feedback spreading control under speed constraints*, SIAM J. Control Optim., 41 (2002), pp. 1281–1294.

[13] K. Kassara, *A set-valued approach to control immunotherapy*, Math. Comput. Modelling, 44 (2006), pp. 1114–1125.

[14] D. Kirschner and J. C. Panetta, *Modelling immunotherapy of the tumor-immune interaction*, J. Math. Biol., 37 (1998), pp. 235–252.

[15] V. A. Kuznetsov and G. D. Knott, *Modeling tumor regrowth and immunotherapy*, Math. Comput. Modelling, 33 (2001), pp. 1275–1287.

[16] U. Ledzewicz and H. Schättler, *Antiangiogenic therapy in cancer treatment as an optimal control problem*, SIAM J. Control Optim., 46 (2007), pp. 1052–1079.

[17] A. Matzavinos, M. A. Chaplain, and J. V. A. Kuznetsov, *Mathematical modelling of the spatio-temporal response of cytotoxic T-lymphocytes to a solid tumor*, Math. Med. Biol., 21 (2004), pp. 1–34.

[18] F. Nani and H. I. Freedman, *A mathematical model of cancer treatment by immunotherapy*, Math. Biosci., 163 (2000), pp. 159–199.

[19] L. G. de Pillis, A. E. Radunskaya, and C. L. Wiseman, *A validated mathematical model of cell-mediated immune response to tumor growth*, Cancer Res., 65 (2005), pp. 7950–7958.

[20] L. Preziosi, *Cancer Modelling and Simulation*, CRC Press, Boca Raton, FL, 2003.

[21] T. Roose, S. J. Chapman, and P. K. Maini, *Mathematical models of avascular tumor growth*, SIAM Rev., 49 (2007), pp. 179–208.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 2, pp. 925–940

© 2009 Society for Industrial and Applied Mathematics

A NECESSARY CONDITION FOR DYNAMIC EQUIVALENCE*

JEAN-BAPTISTE POMET†

Abstract. If two control systems on manifolds of the same dimension are dynamic equivalent, we prove that either they are static equivalent, i.e., equivalent via a classical diffeomorphism, or they are both ruled; for systems of different dimensions, the one of higher dimension must be ruled. A ruled system is one whose equations define at each point in the state manifold a ruled submanifold of the tangent space. Dynamic equivalence is also known as equivalence by endogenous dynamic feedback or by a Lie–Bäcklund transformation when control systems are viewed as underdetermined systems of ordinary differential equations; it is very close to absolute equivalence for Pfaffian systems. It was already known that a differentially flat system must be ruled; this was a particular case of the present result, in which one of the systems was assumed to be "trivial" (or linear controllable).

Key words. control systems, ordinary differential equations, underdetermined systems, dynamic equivalence, absolute equivalence, ruled submanifolds

AMS subject classifications. 34C41, 34L30, 93B17, 93B29

DOI. 10.1137/080723351

1. Introduction.

We consider time-invariant control systems or underdetermined systems of ordinary differential equations (ODEs) where the independent variable is time. *Static equivalence* refers to equivalence via a diffeomorphism in the variables of the equation, or in the state and control variables, with a triangular structure that induces a diffeomorphism (preserving time) in the state variables too. It is also known as "feedback equivalence." *Dynamic equivalence* refers to equivalence via invertible transformations in jet spaces that do not induce any diffeomorphism in a finite number of variables, except when it coincides with static equivalence; these transformations are also known as endogenous dynamic feedback [15, 6], or Lie–Bäcklund transformations [1, 6, 16], although this terminology is more common for systems of partial differential equations; dynamic equivalence is also very close to absolute equivalence for Pfaffian systems [4, 18, 19].

The literature on classification and invariants for static equivalence is too large to be quoted here; let us recall only that, as evidenced by all detailed studies and mentioned in [21], each equivalence class (within control systems on the same manifold or germs of control systems) is very very thin; indeed, it has infinite codimension except in trivial cases. Since dynamic equivalence is a priori more general, it is natural to ask how more general it is. Systems on manifolds of different dimension may be dynamic equivalent but not static equivalent. Restricting our attention to systems on the same manifold and considering dynamic equivalence instead of static, how much bigger are the equivalence classes?

The literature on dynamic feedback linearization [11, 5], differential flatness [6, 15], or absolute equivalence [18] tends to describe the classes containing linear controllable systems or "trivial" systems. The authors of [6, 15, 18] made the link with deep differential geometric questions dating back to [9, 4, 10]; see [2] for a recent overview. Despite these efforts, no characterization is available except for systems with one control, i.e., whose general solution depends on one function of one variable;

*Received by the editors May 6, 2008; accepted for publication (in revised form) November 7, 2008; published electronically February 25, 2009.
http://www.siam.org/journals/sicon/48-2/72335.html

†INRIA, B.P. 93, 06902 Sophia Antipolis cedex, France (Jean-Baptiste.Pomet@sophia.inria.fr).

there are many systems that one suspects to be nonflat, i.e., dynamic equivalent to no trivial system, while no proof is available; see the remark on (23) in section 4.1. There is, however, one powerful necessary condition [17, 20]: a flat system must be ruled; i.e., its equations must define a ruled submanifold in each tangent space. As pointed out in [17], this proves that the equivalence class of linear systems for dynamic equivalence, although bigger than for static equivalence, still has infinite codimension.

Deciding whether two general systems are dynamic equivalent is at least as difficult. There is no method to prove that two systems are not dynamic equivalent. The contribution of this paper is a necessary condition for two systems to be dynamic equivalent that generalizes [17, 20]: if they live on manifolds of the same dimension, either they are static equivalent or they are both ruled; if not, the one of higher dimension must be ruled. Besides being useful to prove that some pairs of systems are not dynamic equivalent, it also implies that "generic" equivalence classes for dynamic equivalence are the same as for static equivalence.

Outline. Notations on jet bundles and differential operators are recalled in section 2; the notions of systems, ruled systems, and dynamic and static equivalence are precisely defined in section 3. Our main result is stated and commented on in section 4 and proved in section 5.

2. Miscellaneous notations.

Let M be an n-dimensional manifold, either $\mathbf{C}^\infty$ (infinitely differentiable) or $\mathbf{C}^\omega$ (real analytic).

2.1. Jet bundles.

Using the notations and definitions of [8, Chapter II, section 2], $J^k(\mathbb{R},M)$ denotes the kth jet bundle of maps $\mathbb{R}\to M$. It is a bundle both over $\mathbb{R}$ and over M. If $(x^1,\dots,x^n)$ is a system of coordinates on an open subset of M, coordinates on the lift of this open subset are given by $t, x^1,\dots,x^n,\dot x^1,\dots,\dot x^n,\dots,$ $(x^1)^{(k)},\dots,(x^n)^{(k)}$, where t is the projection on $\mathbb{R}$.

As an additive group, $\mathbb{R}$ acts on $J^k(\mathbb{R},M)$ by translation of the t-component; the quotient by this action is well defined, and we denote it by

$$J^k(M) \;=\; J^k(\mathbb{R},M)\,/\,\mathbb{R}\;. \tag{1}$$

Since we study only time-invariant systems, we prefer to work with $J^k(M)$. Quotienting indeed drops the t information: local coordinates on $J^k(M)$ are given by $x^1,\dots,x^n,\dot x^1,\dots,\dot x^n,\dots,(x^1)^{(k)},\dots,(x^n)^{(k)}$; for short, we write $x,\dot x,\dots,x^{(k)}$. For $\ell<k$, there is a canonical projection

$$\pi_{k,\ell}:\; J^k(M)\to J^\ell(M) \tag{2}$$

that makes $J^k(M)$ a bundle over $J^\ell(M)$; in particular, it is a bundle over $M=J^0(M)$ and over $\mathrm{T}M=J^1(M)$. In coordinates,

$$\pi_{k,\ell}\left(x,\dot x,\dots,x^{(\ell)},\dots,x^{(k)}\right)=\left(x,\dot x,\dots,x^{(\ell)}\right)\;.$$

Notation. To a subset $\Omega\subset J^k(M)$, we associate, for all ℓ, a subset $\Omega_\ell\subset J^\ell(M)$ in the following manner (obviously, $\Omega_k=\Omega$):

$$\Omega_\ell=\begin{cases}\pi_{k,\ell}(\Omega) & \text{if } \ell\le k,\\ {\pi_{\ell,k}}^{-1}(\Omega) & \text{if } \ell\ge k.\end{cases} \tag{3}$$

2.2. The kth jet of a smooth ($\mathbf{C}^\infty$) map $x(.) : I \longrightarrow M$. With $I \subset \mathbb{R}$ a time interval, it is a smooth map $j^k(\,x(.)\,) : I \to J^k(M)$ (see again [8]); in coordinates,

$$j^k(\,x(.)\,)\,(t) = \Big(x(t), \dot{x}(t), \ddot{x}(t), \dots, x^{(k)}(t)\Big) \ .$$

By *a smooth map whose kth jet remains in* Ω, for some $\Omega \subset J^k(M)$, we mean a smooth $x(.) : I \to M$ such that $j^k(\,x(.)\,)(t) \in \Omega$ for all t in I.

2.3. Differential operators. If Ω is an open subset of $J^k(M)$ and M' is a manifold of dimension n', a smooth ($\mathbf{C}^\infty$ or $\mathbf{C}^\omega$) map $\Phi : \Omega \to M'$ defines the smooth differential operator of order[1] k

$$\mathcal{D}^k_\Phi \ = \ \Phi \circ j^k \ . \tag{4}$$

Obviously, $\mathcal{D}^k_\Phi$ sends smooth maps $I \to M$ whose kth jet remains in Ω to smooth maps $I \to M'$. In coordinates, the image of $t \mapsto x(t)$ is $t \mapsto \Phi(x(t), \dot{x}(t), \ddot{x}(t), \dots, x^{(k)}(t))$. Note that we do not require that k be minimal, so Φ *might* not depend on $x^{(k)}$.

We call $j^r \circ \mathcal{D}^k_\Phi$ the *rth prolongation of* the differential operator $\mathcal{D}^k_\Phi$; it sends smooth maps $I \to M$ whose kth jet remains in Ω to smooth maps $I \to J^r(M')$; it is indeed the differential operator $\mathcal{D}^{k+r}_{\Phi^{[r]}}$, of order $k + r$, with $\Phi^{[r]}$ the unique smooth map ${\pi_{k+r,k}}^{-1}(\Omega) \to J^r(M')$ such that

$$j^r \circ \Phi \circ j^k \ = \ \Phi^{[r]} \circ j^{k+r} \ . \tag{5}$$

We call $\Phi^{[r]}$ the rth prolongation of Φ. One has $\pi_{r,0} \circ \Phi^{[r]} = \Phi \circ \pi_{k+r,k}$ and more generally, for $s < r$,

$$\pi_{r,s} \circ \Phi^{[r]} = \Phi^{[s]} \circ \pi_{k+r,k+s} \ . \tag{6}$$

3. Systems and equivalence.

3.1. Systems.

DEFINITION 3.1. *A* $\mathbf{C}^\infty$ *or* $\mathbf{C}^\omega$ regular system *with m controls on a smooth manifold M is a* $\mathbf{C}^\infty$ *or* $\mathbf{C}^\omega$ *subbundle Σ of the tangent bundle* $\mathrm{T}M$

$$\begin{tikzcd}
\Sigma \arrow[r, hook, "i"] \arrow[dr, "\pi"'] & \mathrm{T}M \arrow[d] \\
& M
\end{tikzcd} \tag{7}$$

with fiber Υ, a $\mathbf{C}^\infty$ *or* $\mathbf{C}^\omega$ *manifold of dimension m (e.g., an open subset of $\mathbb{R}^m$). The* velocity set *at a point $x \in M$ is the fiber $\Sigma_x = \pi^{-1}(\{x\})$, a submanifold of $\mathrm{T}_x M$ diffeomorphic to Υ.*

DEFINITION 3.2 (solutions of a system). *A solution of system Σ on the real interval I is a* smooth *(*$\mathbf{C}^\infty$*) $x(.) : I \to M$ such that $j^1(x(.))(t) \in \Sigma$ for all $t \in I$.*

Although a general solution of a system need not be smooth, we consider only *smooth* solutions. They form a rich enough class in the sense that systems are fully characterized by their set of smooth solutions.

Locally, one may write "explicit" equations of Σ in the following form. Of course, there are many choices of coordinates, and the map f depends on this choice.

PROPOSITION 3.3. *For each $\xi \in \Sigma$, with $\Sigma \hookrightarrow \mathrm{T}M$ a regular system* (7), *there are*

[1] "Of order no larger than k" would be more accurate: if Φ does not depend on kth derivatives, the order in the usual sense would be smaller than k. See, for instance, Ψ in example (22).

- *an open neighborhood $\mathcal{U}$ of ξ in $\mathrm{T}M$, $\mathcal{U}_0$ its projection on M,*
- *a system of local coordinates $(x_{\mathrm{I}}, x_{\mathbb{I}})$ on $\mathcal{U}_0$, with x_{I} a block of dimension $n-m$ and $x_{\mathbb{I}}$ of dimension m,*
- *an open subset U of $\mathbb{R}^{n+m}$ and a smooth ($\mathbf{C}^\infty$ or $\mathbf{C}^\omega$) map $f : U \to \mathbb{R}^{n-m}$*

such that the equation of $\Sigma \cap \mathcal{U}$ in these coordinates is

$$\dot{x}_{\mathrm{I}} = f(x_{\mathrm{I}}, x_{\mathbb{I}}, \dot{x}_{\mathbb{I}}), \quad (x_{\mathrm{I}}, x_{\mathbb{I}}, \dot{x}_{\mathbb{I}}) \in U \ . \tag{8}$$

Proof. The proof is a consequence of the implicit function theorem. □

Control systems. A more usual representation of a system with m controls is

$$\dot{x} = F(x, u) \ , \quad x \in M \ , \ u \in \mathcal{B} \ , \tag{9}$$

with $\mathcal{B}$ an open subset of $\mathbb{R}^m$ and $F : M \times \mathcal{B} \to \mathrm{T}M$ smooth enough. It can be brought locally, in block coordinates $(x_{\mathrm{I}}, x_{\mathbb{I}})$, to the form

$$\dot{x}_{\mathrm{I}} = f(x_{\mathrm{I}}, x_{\mathbb{I}}, u) \ , \quad \dot{x}_{\mathbb{I}} = u, \tag{10}$$

modulo a static feedback on u, at least around nonsingular points (x, u) where

$$\operatorname{rank} \frac{\partial F}{\partial u}(x, u) = m \ . \tag{11}$$

Equation (8) can be obtained by eliminating the control u in (10).

If (11) holds, (9) defines a system in the sense of Definition 3.1. All results on systems in that sense may easily be translated to control systems (9).

Implicit systems of ODEs. A smooth *system of $n - m$ ODEs* on M: $R(x, \dot{x}) = 0$ with $R : \mathrm{T}M \to \mathbb{R}^{n-m}$ also defines a system in the sense of Definition 3.1 if it is nonsingular, i.e., rank $\frac{\partial R}{\partial \dot{x}}(x, \dot{x}) = n - m$.

Singularities. With the above rank assumptions or the one that Σ is a subbundle in Definition 3.1, we carefully avoid singular systems. This paper does not apply to singular control systems or singular implicit systems of ODEs.

Prolongations of Σ. For integers $k \geq 1$, we denote by Σ_k the prolongation of the system Σ to kth order; it is *the* subbundle $\Sigma_k \hookrightarrow J^k(M)$ with the following property: for any smooth map $x(.) : I \to M$, with $j^k(x(.))$ defined in section 2.2,

$$j^1(x(.))(t) \in \Sigma \ , \ t \in I \quad \Leftrightarrow \quad j^k(x(.))(t) \in \Sigma_k \ , \ t \in I \ . \tag{12}$$

The left-hand side means that $x(.)$ is a *solution* of Σ according to Definition 3.2. Obviously, $\Sigma_1 = \Sigma$. We may describe Σ_k in coordinates.

Proposition 3.4. *Let K be a positive integer. There is a unique subbundle $\Sigma_K \hookrightarrow J^K(M)$ such that*

(13) *a smooth map $x(.) : I \to M$ is a solution of system Σ on the real interval I if and only if $j^K(x(.))(t) \in \Sigma_K$ for all $t \in I$.*

For all $\xi \in \Sigma_K$, its projection $\xi_1 = \pi_{K,1}(\xi)$ is in Σ and, with $\mathcal{U}$ the neighborhood of ξ_1, $(x_{\mathrm{I}}, x_{\mathbb{I}})$ the coordinates on $\mathcal{U}_0$, U the open subset $\mathbb{R}^{n+m}$, and $f : U \to \mathbb{R}^m$ the map given by Proposition 3.3, *the equations of $\mathcal{U}_K \cap \Sigma_K$ in $J^K(M)$ are, in the coordinates $(x_{\mathrm{I}}, x_{\mathbb{I}}, \dot{x}_{\mathrm{I}}, \dot{x}_{\mathbb{I}}, \ldots, x_{\mathrm{I}}^{(K)}, x_{\mathbb{I}}^{(K)})$ induced on $\mathcal{U}$ by $(x_{\mathrm{I}}, x_{\mathbb{I}})$,*

$$\begin{aligned} & x_{\mathrm{I}}^{(i)} = f^{(i-1)}\left(x_{\mathrm{I}}, x_{\mathbb{I}}, \dot{x}_{\mathbb{I}}, \ldots, x_{\mathbb{I}}^{(i)}\right) \ , \quad 1 \leq i \leq K \ , \\ & \left(x_{\mathrm{I}}, x_{\mathbb{I}}, \dot{x}_{\mathbb{I}}, \ldots, x_{\mathbb{I}}^{(K)}\right) \ \in \ U \times \mathbb{R}^{(K-1)m} \ , \end{aligned} \tag{14}$$

where, for a smooth map $f : U \to \mathbb{R}^{n-m}$ *and* $\ell \geq 0$, $f^{(\ell)}$ *is the smooth map* $U \times \mathbb{R}^{Km} \to \mathbb{R}^{n-m}$ *defined by* $f^{(0)} = f$ *and, for* $i \geq 1$,

$$f^{(i)}\left(x_{\mathrm{I}}, x_{\mathbb{I}}, \dot{x}_{\mathbb{I}}, \ldots, x_{\mathbb{I}}^{(i+1)}\right) = \frac{\partial f^{(i-1)}}{\partial x_{\mathrm{I}}} f\left(x_{\mathrm{I}}, x_{\mathbb{I}}, \dot{x}_{\mathbb{I}}\right) + \sum_{i=0}^{i} \frac{\partial f^{(i-1)}}{\partial x_{\mathbb{I}}^{(i)}} x_{\mathbb{I}}^{(i+1)} \ . \tag{15}$$

Proof. The proof is classical and obvious in coordinates. □

Remark 3.5. Each Σ_{k+1} $(k \geq 1)$ is an affine bundle over Σ_k and may be viewed as an affine subbundle of $\mathrm{T}\Sigma_k$; i.e., it is a system in the sense of section 3.1 on the manifold Σ_k instead of M.

In particular, $\Sigma_2 \hookrightarrow \mathrm{T}\Sigma$ is the system obtained by “adding an integrator in each control” of the system $\Sigma \hookrightarrow \mathrm{T}M$. It is an affine system (i.e., affine subbundle) even when Σ is not.

3.2. Ruled systems. Recall that a smooth submanifold of an affine space is ruled if and only if it is a union of straight lines, i.e., if through each point of the submanifold passes a straight line contained in the submanifold. Such a manifold must be unbounded; since we want to consider the intersection of a submanifold with an arbitrary open set and allow this patch to be “ruled,” we use the same slightly abusive notion as [14]: a submanifold N is ruled if and only if, through each point of it, passes a straight line which is contained in N “until it reaches the boundary of N.” Here, the boundary of the submanifold N is $\partial N = \overline{N} \setminus N$.

A system will be called ruled if and only if Σ_x is, for all x, a ruled submanifold of $\mathrm{T}_x M$. This is formalized below in a self-contained manner.

Definition 3.6. *Let* $\mathcal{O}$ *be an open subset of* $\mathrm{T}M$. *System* Σ *(see* (7)*) is* ruled in $\mathcal{O}$ *if and only if, for all* $(x, \dot{x}) \in (\mathcal{O} \cap \Sigma)$, *there are a nonzero vector* $w \in \mathrm{T}_x M \setminus \{0\}$ *and two possibly infinite numbers* $\lambda^- \in [-\infty, 0)$ *and* $\lambda^+ \in (0, +\infty]$ *such that* $(x, \dot{x} + \lambda w) \in \mathcal{O} \cap \Sigma$ *for all* λ, $\lambda^- < \lambda < \lambda^+$ *and*

$$\begin{aligned} \lambda^- > -\infty &\Rightarrow (x, \dot{x} + \lambda^- w) \in \partial\left(\mathcal{O} \cap \Sigma\right) , \\ \lambda^+ < +\infty &\Rightarrow (x, \dot{x} + \lambda^+ w) \in \partial\left(\mathcal{O} \cap \Sigma\right) . \end{aligned} \tag{16}$$

Recall that, by definition, $\partial\left(\mathcal{O} \cap \Sigma\right) = \overline{\mathcal{O} \cap \Sigma} \setminus (\mathcal{O} \cap \Sigma)$.

We shall need the following characterization.

Proposition 3.7 (see [14]). *Let* $\mathcal{O}$ *be an open subset of* $\mathrm{T}M$. Σ *is ruled in* $\mathcal{O}$ *if and only if, for all* $\xi = (x, \dot{x})$ *in* $\Sigma \cap \mathcal{O}$, *there is a straight line in* $\mathrm{T}_x M$ *passing through* $\dot{x}$ *that has contact of infinite order with* Σ_x *at* $\dot{x}$.

Proof. From [14, Theorem 1], a “patch of” submanifold of dimension m in a manifold of dimension n is ruled if and only if there is, through each point, a straight line that has contact of order $n + 1$. This is of course implied by infinite order. □

3.3. Dynamic equivalence. The following notion is usually called dynamic equivalence or equivalence by (endogenous) dynamic feedback transformations in control theory; see [15, 7, 12, 16]. It is in fact also the notion of a Lie–Bäcklund transformation, limited to ordinary differential equations, as noted in [7] or [16].

Definition 3.8. *Let* $k = \infty$ *or* $k = \omega$. *Let* $\Sigma \hookrightarrow \mathrm{T}M$ *and* $\Sigma' \hookrightarrow \mathrm{T}M'$ *be* $\mathbf{C}^k$ *regular systems (see* (7)*) on two manifolds* M *and* M', K, K' *two integers, and* $\Omega \subset J^K(M)$ *and* $\Omega' \subset J^{K'}(M')$ *two open subsets.*

Systems Σ *and* Σ' *are* dynamic equivalent over Ω and Ω' *if and only if there exists two mappings of class* $\mathbf{C}^k$:

$$\Phi : \Omega \to M' , \qquad \Psi : \Omega' \to M \tag{17}$$

inducing differential operators $\mathcal{D}^K_\Phi$ *and* $\mathcal{D}^{K'}_\Psi$ *(see* (4)*) such that, for any interval* I*,*

- *for any solution* $x(.) : I \to M$ *of* Σ *whose* K*th jet remains inside* Ω*,* $\mathcal{D}^K_\Phi(\,x(.)\,)$ *is a solution of* Σ' *whose* K'*th jet remains inside* Ω' *and* $\mathcal{D}^{K'}_\Psi(\,\mathcal{D}^K_\Phi(\,x(.)\,)\,) \;=\; x(.)$*;*
- *for any solution* $z(.) : I \to M'$ *of* Σ' *whose* K'*th jet remains inside* Ω'*,* $\mathcal{D}^{K'}_\Psi(\,z(.)\,)$ *is a solution of* Σ *whose* K*th jet remains inside* Ω *and* $\mathcal{D}^K_\Phi(\,\mathcal{D}^{K'}_\Psi(\,z(.)\,)\,) \;=\; z(.)$*.*

Remark 3.9. Since all properties are tested on *solutions*, only the restriction of Φ and Ψ to Σ_K and $\Sigma_{K'}$ (see Proposition 3.4) matter; for instance, Φ can be arbitrarily modified away from Σ_K without changing any conclusions. Borrowing this language from the literature on Lie–Bäcklund transformations, Φ and Ψ above are "external" correspondences.

In [7] or in [16], the "internal" point of view prevails: for instance, Φ and Ψ are replaced, in [7], by diffeomorphisms between diffieties. This is more intrinsic because maps are defined only where they are to be used. However, the definitions are equivalent because these internal maps admit infinitely many "external" prolongations.

Here, this external point of view is adopted because it makes the statement of the main result less technical. Note, however, that, as a preliminary to the proofs, an "internal" translation is given in section 5.1.

Remark 3.10. In the theorems, we shall require that Ω and Ω' satisfy

$$\Omega_1 \cap \Sigma \;\subset\; (\Omega \cap \Sigma_K)_1 \quad\text{and}\quad \Omega'_1 \cap \Sigma' \;\subset\; (\Omega' \cap {\Sigma'}_{K'})_1\,; \tag{18}$$

i.e., any (jet of) solution whose first jet is in Ω_1 lifts to at least one (jet of) solution whose Kth jet is in Ω. Note the following facts about this requirement.

- These inclusions are equalities, for the reverse inclusions always hold.
- Replacing the original Ω with $\Omega\backslash(\overline{(\Omega_1 \cap \Sigma) \setminus (\Omega \cap \Sigma_K)_1})_K$ and Ω' accordingly forces (18); alternatively, keeping arbitrary open sets, Theorems 4.2 and 4.1 would hold with Ω_1 replaced with $\Omega_1 \setminus \overline{(\Omega_1 \cap \Sigma) \setminus (\Omega \cap \Sigma_K)_1}$.
- When $\Sigma' = \mathrm{T}M'$ is the trivial system (see section 3.5), any open Ω' satisfies (18).

3.4. Static equivalence.

Definition 3.11. *Let* $\mathcal{O} \subset \mathrm{T}M$ *and* $\mathcal{O}' \subset \mathrm{T}M'$ *be open subsets. Systems* Σ *and* Σ' *are* static equivalent *over* $\mathcal{O}$ *and* $\mathcal{O}'$ *if and only if there is a smooth diffeomorphism* $\Phi : \mathcal{O}_0 \to \mathcal{O}'_0$ *such that the following holds:*

$$\left.\begin{array}{l}\textit{a smooth map } t \mapsto x(t) \textit{ is a solution of } \Sigma \textit{ whose first jet remains in } \mathcal{O}\\ \textit{if and only if } t \mapsto \Phi(x(t)) \textit{ is a solution of } \Sigma' \textit{ whose first jet remains in } \mathcal{O}'.\end{array}\right\} \tag{19}$$

Definition 3.12 (local static equivalence). *Let* $\mathcal{O} \subset \mathrm{T}M$ *and* $\mathcal{O}' \subset \mathrm{T}M'$ *be open subsets. Systems* Σ *and* Σ' *are* locally *static equivalent over* $\mathcal{O}$ *and* $\mathcal{O}'$ *if and only if there are coverings of* $\mathcal{O} \cap \Sigma$ *and* $\mathcal{O}' \cap \Sigma'$ *:*

$$\Sigma \cap \mathcal{O} \;\subset\; \Sigma \cap \bigcup_{\alpha \in A} \mathcal{O}^\alpha\,, \quad \Sigma' \cap \mathcal{O}' \;\subset\; \Sigma' \cap \bigcup_{\alpha \in A} \mathcal{O}'^\alpha,$$

where A *is a set of indices and* $\mathcal{O}^\alpha$ *and* $\mathcal{O}'^\alpha$ *are open subsets of* $\mathcal{O}$ *and* $\mathcal{O}'$ *such that, for all* α*, systems* Σ *and* Σ' *are static equivalent over* $\mathcal{O}^\alpha$ *and* $\mathcal{O}'^\alpha$*.*

This definition, stated in terms of solutions, is translated into point (a) below that relies only on the geometry of Σ and Σ' as submanifolds. Point (b) is used, for instance, in [13, 22] where the "centro-affine" geometry of each Σ_x is studied.

PROPOSITION 3.13. (a) *Systems* Σ *and* Σ' *are* static equivalent *over* $\mathcal{O} \subset \mathrm{T}M$ *and* $\mathcal{O}' \subset \mathrm{T}M'$ *if and only if there is a smooth diffeomorphism* $\Phi : \mathcal{O}_0 \to \mathcal{O}'_0$ *such that* $\Phi_\star$ *maps* $\mathcal{O} \cap \Sigma$ *to* $\mathcal{O}' \cap \Sigma'$.

(b) *If systems* Σ *and* Σ' *are static equivalent over* $\mathcal{O} \subset \mathrm{T}M$ *and* $\mathcal{O}' \subset \mathrm{T}M'$, *there is, for each* $x \in \mathcal{O}_0$, *a* linear isomorphism $\mathrm{T}_x M \to \mathrm{T}_{\Phi(x)} M'$ *that maps* Σ_x *to* $\Sigma'_{\Phi(x)}$.

(c) *Static equivalence preserves ruled systems.*

Proof. (b) and (c) are easy consequences of (a), which in turn is clear by differentiating solutions in Definition 3.2. □

3.5. Examples. 1. We call *trivial system* on a smooth manifold M the tangent bundle itself $\mathrm{T}M$. Any smooth $x(.) : I \to M$ is a solution of this system; it corresponds to "no equation," to the control system $\dot{x} = u$, or to the "affine diffieties" in [7]. Following [6, 7], a system $\Sigma \hookrightarrow \mathrm{T}M$ is called *differentially flat* (on $\Omega \subset J^K(M)$) if and only if it is dynamic equivalent (over Ω and Ω') to the trivial system $\mathrm{T}M'$ for some manifold M'.

2. Any system $\Sigma \hookrightarrow \mathrm{T}M$ is dynamic equivalent to the one obtained by "adding integrators." It was described in Remark 3.5 as an affine subbundle $\Sigma_2 \hookrightarrow \mathrm{T}\Sigma$; Σ and Σ_2 are equivalent in the sense of Definition 3.8 with $M' = \Sigma$, $K = 1$, $K' = 0$, and Ω an open neighborhood of Σ in $J^1(M) = \mathrm{T}M$ such that there is a $\Phi : \Omega \to \Sigma$ that coincides with identity on Σ, $\Omega' = M' = \Sigma$, and $\Psi = \pi$ (see (7)).

This may be easier to follow in the coordinates of Proposition 3.3. The prolongation of (8) has state $(y_{\mathrm{I}}, y_{\mathbb{I}}) \in U$, with y_{I} a block of dimension n and $y_{\mathbb{I}}$ of dimension m, and equation $\dot{y}_{\mathrm{I}} = (f(y_{\mathrm{I}}, y_{\mathbb{I}})\,,\, y_{\mathbb{I}})$. In coordinates, the transformations $\Phi : J^1(U_0) \to U$ and $\Psi : U \to U_0$ are given by $(y_{\mathrm{I}}, y_{\mathbb{I}}) = \Phi(x_{\mathrm{I}}, x_{\mathbb{I}}, \dot{x}_{\mathrm{I}}, \dot{x}_{\mathbb{I}}) = (x, \dot{x}_{\mathbb{I}})$ and $x = \Psi(y) = y_{\mathrm{I}}$, respectively.

Static equivalence between these systems of different dimension does not hold.

3. Let us now give, mostly to illustrate the role of the integers K, K' and the open sets Ω and Ω', two more specific examples of systems $\Sigma \hookrightarrow \mathrm{T}\mathbb{R}^3$ and $\Sigma' \hookrightarrow \mathrm{T}\mathbb{R}^3$ with the following equations in $\mathrm{T}\mathbb{R}^3$, with coordinates $(x_1, x_2, x_3, \dot{x}_1, \dot{x}_2, \dot{x}_3)$ or $(y_1, y_2, y_3, \dot{y}_1, \dot{y}_2, \dot{y}_3)$, clearly defining subbundles with fiber diffeomorphic to $\mathbb{R}^2$:

$$\Sigma : \ \dot{x}_1 = x_2\ , \qquad \Sigma' : \ \dot{y}_1 = y_2 + (\dot{y}_2 - y_1 \dot{y}_3)\, \dot{y}_3\ . \tag{20}$$

These equations are even globally in the "explicit" form given by Proposition 3.3.

First of all, Σ is dynamic equivalent to the trivial system $\Sigma'' = \mathrm{T}\mathbb{R}^2$, with $\Phi : \mathbb{R}^3 \to \mathbb{R}^2$ defined by $\Phi(x_1, x_2, x_3) = (x_1, x_3)$ and $\Psi : J^1(\mathbb{R}^2) \to \mathbb{R}^3$ given by $\Psi(z_1, z_2, \dot{z}_1, \dot{z}_2) = (z_1, \dot{z}_1, z_2)$. Here, $K = 0, K' = 1, \Omega = \mathbb{R}^2$, and $\Omega' = J^1(\mathbb{R}^2)$.

Also, with $K = 1$ and $K' = 2$, systems Σ *and* Σ' *are dynamic equivalent* over $\Omega \subset J^1(\mathbb{R}^3)$ and $\Omega' \subset J^2(\mathbb{R}^3)$ defined by

$$\begin{aligned} \Omega &= \left\{(x_1, x_2, x_3,\, \dot{x}_1, \dot{x}_2, \dot{x}_3),\ 1 - \dot{x}_2 - {x_2}^3 \neq 0\right\}\ , \\ \Omega' &= \left\{(y_1, y_2, y_3,\, \dot{y}_1, \dot{y}_2, \dot{y}_3,\, \ddot{y}_1, \ddot{y}_2, \ddot{y}_3),\ 1 - \ddot{y}_3 - \dot{y}_3^3 \neq 0\right\}\ . \end{aligned}$$

The maps $\Phi : \Omega \to \mathbb{R}^3$ and $\Psi : \Omega' \to \mathbb{R}^3$ are given by

$$\Phi(x_1, x_2, x_3,\, \dot{x}_1, \dot{x}_2, \dot{x}_3) = \left(\frac{(1 - \dot{x}_2) x_3 + x_2\, \dot{x}_3}{1 - \dot{x}_2 - {x_2}^3}\,,\, \frac{{x_2}^2\, x_3 + \dot{x}_3}{1 - \dot{x}_2 - {x_2}^3}\,,\, x_1 \right)\,, \tag{21}$$

$$\Psi(y_1, y_2, y_3,\, \dot{y}_1, \dot{y}_2, \dot{y}_3,\, \ddot{y}_1, \ddot{y}_2, \ddot{y}_3) = (\, y_3\,,\, \dot{y}_3\,,\, y_1 - \dot{y}_3\, y_2\,)\,. \tag{22}$$

Remark 3.14. Since Ψ does not depend on second derivatives, $K' = 2$ is not the order of the differential operator $\mathcal{D}_\Psi^{K'}$ in the usual sense; this illustrates the footnote after (4); it is, however, necessary to go to second jets to describe the domain Ω' where the restriction to solutions of Σ' of this first order operator can be inverted.

Finally, note that systems Σ *and* Σ' *are not static equivalent* because, from Proposition 3.13(b), this would imply that each Σ_x is sent to some Σ'_y by a linear isomorphism $\mathrm{T}_xM \to \mathrm{T}_yM'$, which is not possible because each Σ_x is an affine subspace of T_xM and Σ'_y a nondegenerate quadric of T_yM'.

4. Consider two more systems $\Sigma \hookrightarrow \mathrm{T}\mathbb{R}^3$ and $\Sigma' \hookrightarrow \mathrm{T}\mathbb{R}^3$ described as in (20):

$$(23) \qquad \Sigma :\ \dot{x}_1 = x_2 + (\dot{x}_2 - x_1\dot{x}_3)^2\,\dot{x}_3^{\,2}\ , \qquad \Sigma' :\ \dot{y}_1 = y_2 + (\dot{y}_2 - y_1\dot{y}_3)^2\,\dot{y}_3\ .$$

System Σ is ruled (each Σ_y is the union of lines $\dot{y}_2 - y_1\dot{y}_3 = \lambda$, $\dot{y}_1 = y_2 + \lambda^2\,\dot{y}_3$ for λ in $\mathbb{R}$), while Σ' is not. Hence, from point (c) of Proposition 3.13, Σ *and* Σ' *are not static equivalent.* We shall come back to these two systems from the point of view of flatness and dynamic equivalence in sections 4.1 and 4.3.

4. Necessary conditions.

4.1. The case of flatness. It has been known since [17, 20] that a system which is dynamic equivalent to a *trivial system* (see the beginning of section 3.5; such a system is called differentially flat) must be ruled; of course, at least in the smooth case, this is true only on the domain where equivalence is assumed.

THEOREM 4.1 (see [17, 20]). *If* Σ *is dynamic equivalent to the trivial system* $\Sigma'{=}\mathrm{T}M'$ *over* $\Omega \subset J^K(M)$ *and* $\Omega' \subset J^{K'}(M')$ *satisfying* (18), *then* Σ *is ruled in* Ω_1.

Application. Since Σ in (23) is not ruled, this theorem implies that it is not flat, i.e., not dynamic equivalent to the trivial system $\mathrm{T}\mathbb{R}^2$. On the contrary, Σ' in (23) is ruled, and, hence, the result does not help to decide if it is flat or not; in fact, one conjectures that this system is not flat, but no proof is available; see [3].

4.2. Main idea of the proofs. Our main result, stated in the next section, studies what remains of Theorem 4.1 when Σ' is not the trivial system. Due to many technicalities concerning regularity conditions, the main ideas may be difficult to grasp in the proof given in section 5.2. In order to enlighten these ideas and even the result itself, let us first sketch the proof of the above theorem following the line of [17] (itself inspired from [10]), but *without assuming a priori that* Σ' *is trivial.*

Take two arbitrary systems Σ and Σ', and assume that they are dynamic equivalent. From Proposition 3.3, one may use locally the explicit forms

$$\Sigma :\ \dot{x}_{\mathrm{I}} = f(x_{\mathrm{I}}, x_{\mathbb{I}}, \dot{x}_{\mathbb{I}})\ , \quad \Sigma' :\ \dot{z}_{\mathrm{I}} = g(z_{\mathrm{I}}, z_{\mathbb{I}}, \dot{z}_{\mathbb{I}})\ .$$

Recall that n and n' denote the dimensions of x and z; assume $n \le n'$. Since we work only on *solutions* (see Remark 3.9 and also section 5.1) and the above equations allow one to express each time derivative $x_{\mathrm{I}}^{(j)}$, $j \ge 1$, as a function of $x_{\mathrm{I}}, x_{\mathbb{I}}$, $\dot{x}_{\mathbb{I}}$, $\dots$, $x_{\mathbb{I}}^{(j)}$, we may work with the variables $x_{\mathrm{I}}, x_{\mathbb{I}}, \dot{x}_{\mathbb{I}}, \ddot{x}_{\mathbb{I}}, x_{\mathbb{I}}^{(3)}, \dots$ and $z_{\mathrm{I}}, z_{\mathbb{I}}, \dot{z}_{\mathbb{I}}, \ddot{z}_{\mathbb{I}}, z_{\mathbb{I}}^{(3)}, \dots$ only. The map Φ of Definition 3.8 translates, in these coordinates, into a correspondence $z_{\mathrm{I}} = \phi_{\mathrm{I}}(x_{\mathrm{I}}, x_{\mathbb{I}}, \dot{x}_{\mathbb{I}}, \dots, x_{\mathbb{I}}^{(K)})$, $z_{\mathbb{I}} = \phi_{\mathbb{I}}(x_{\mathrm{I}}, x_{\mathbb{I}}, \dot{x}_{\mathbb{I}}, \dots, x_{\mathbb{I}}^{(K)})$; here, the number K is chosen such that *the dependence of* ϕ *versus* $x_{\mathbb{I}}^{(K)}$ *is effective.*

If $K = 0$, this reads $z = \phi(x)$, and $n < n'$ is absurd because it would imply (around points where the rank of ϕ is constant) some nontrivial relations $R(z) = 0$. Hence, $n = n'$, ϕ is a local diffeomorphism, and static equivalence holds locally.

If $K \ge 1$, note that Φ mapping solutions of Σ to solutions of Σ' implies (plug the expression of z given by ϕ into state equations of Σ') the following identity, valid for all $x_{\mathrm{I}}, x_{\mathbb{I}}, \dot{x}_{\mathbb{I}}, \dots, x_{\mathbb{I}}^{(K+1)}$:

$$\begin{aligned}&\frac{\partial \phi_{\mathrm{I}}}{\partial x_{\mathrm{I}}} f(x_{\mathrm{I}}, x_{\mathbb{I}}, \dot{x}_{\mathbb{I}}) + \frac{\partial \phi_{\mathrm{I}}}{\partial x_{\mathbb{I}}} \dot{x}_{\mathbb{I}} + \frac{\partial \phi_{\mathrm{I}}}{\partial \dot{x}_{\mathbb{I}}} \ddot{x}_{\mathbb{I}} + \cdots + \frac{\partial \phi_{\mathrm{I}}}{\partial x_{\mathbb{I}}^{(K)}} x_{\mathbb{I}}^{(K+1)} \\ &\quad = \; g\left(\phi_{\mathrm{I}}, \phi_{\mathbb{I}},\, \frac{\partial \phi_{\mathbb{I}}}{\partial x_{\mathrm{I}}} f(x_{\mathrm{I}}, x_{\mathbb{I}}, \dot{x}_{\mathbb{I}}) + \frac{\partial \phi_{\mathbb{I}}}{\partial x_{\mathbb{I}}} \dot{x}_{\mathbb{I}} + \frac{\partial \phi_{\mathbb{I}}}{\partial x_{\mathbb{I}}} \ddot{x}_{\mathbb{I}} + \cdots + \frac{\partial \phi_{\mathbb{I}}}{\partial x_{\mathbb{I}}^{(K)}} x_{\mathbb{I}}^{(K+1)} \right),\end{aligned}$$

where ϕ_{I} and $\phi_{\mathbb{I}}$ depend on $x_{\mathrm{I}}, x_{\mathbb{I}}, \dot{x}_{\mathbb{I}}, \dots, x_{\mathbb{I}}^{(K)}$ only and, at least at generic points, $\left(\frac{\partial \phi_{\mathrm{I}}}{\partial x_{\mathbb{I}}^{(K)}}, \frac{\partial \phi_{\mathbb{I}}}{\partial x_{\mathbb{I}}^{(K)}} \right) \neq (0,0)$. Fixing such $x_{\mathrm{I}}, x_{\mathbb{I}}, \dot{x}_{\mathbb{I}}, \dots, x_{\mathbb{I}}^{(K)}$ and, consequently, $z = \phi(x_{\mathrm{I}}, x_{\mathbb{I}}, \dot{x}_{\mathbb{I}}, \dots, x_{\mathbb{I}}^{(K)})$ and examining Σ'_z as a submanifold of T_zM' with equation $\dot{z}_{\mathrm{I}} = g(z, \dot{z}_{\mathbb{I}})$, it is clear that moving $x_{\mathbb{I}}^{(K+1)}$ in a direction which is not in the kernel of $\frac{\partial \phi_{\mathbb{I}}}{\partial x_{\mathbb{I}}^{(K)}}(x_{\mathrm{I}}, x_{\mathbb{I}}, \dot{x}_{\mathbb{I}}, \dots, x_{\mathbb{I}}^{(K)})$ provides a straight line of T_zM' contained in Σ'_z and, since this covers all points of Σ'_z, proves that the latter is a ruled submanifold of T_zM' and finally that system Σ' is ruled. We examined only regular points; see section 5.2 for a proper proof.

Collecting the two cases, we have proved that, if $n \leq n'$, either Σ' is ruled or $n = n'$ and Σ' is static equivalent to Σ. This is stated formally in Theorem 4.2.

4.3. The result for general systems. The contribution of this paper is the following strong necessary condition for dynamic equivalence between two general systems. Ω_1 and Ω'_1 are defined by (3).

THEOREM 4.2. *Let Σ and Σ' be systems on manifolds of dimension n and n', K, K' two integers, and $\Omega \subset J^K(M)$, $\Omega' \subset J^{K'}(M')$ two open subsets satisfying* (18).

If Σ and Σ' are dynamic equivalent over Ω and Ω', then

if $n > n'$, *system Σ is ruled in Ω_1;*

if $n < n'$, *system Σ' is ruled in Ω'_1;*

if $n = n'$, *then* *(see Definition* 3.12 *for "locally static equivalent")*

- in the real analytic case *and if $\Omega_1 \cap \Sigma$ and $\Omega'_1 \cap \Sigma'$ are connected, either systems Σ and Σ' are ruled in Ω_1 and Ω'_1, respectively, or they are locally static equivalent over Ω_1 and Ω'_1;*
- in the smooth ($\mathbf{C}^\infty$) case, *there are open subsets $\mathcal{R}, \mathcal{S}$ of Ω_1 and $\mathcal{R}', \mathcal{S}'$ of Ω'_1 such that Ω_1 and Ω'_1 are covered as*

$$\Omega_1 = \overline{\mathcal{R}} \cup \mathcal{S} = \mathcal{R} \cup \overline{\mathcal{S}}, \quad \Omega'_1 = \overline{\mathcal{R}'} \cup \mathcal{S}' = \mathcal{R}' \cup \overline{\mathcal{S}'} \tag{24}$$

and the systems have the following properties on these sets:

1. *Σ and Σ' are ruled in $\mathcal{R}$ and $\mathcal{R}'$, respectively;*
2. *Σ and Σ' are* locally *static equivalent over $\mathcal{S}$ and $\mathcal{S}'$.*

Proof. See section 5.2. □

A few remarks are in order:

1. Theorem 4.1 is a consequence. Indeed, $n' = m'$ because Σ' is trivial, dynamic equivalence implies $m' = m$ (this is common knowledge; see [4], [7] or [16, Theorem 1]), and $n \geq m$ for any system; hence, $n \geq n'$ and Theorem 4.2 directly implies that Σ is ruled except if the systems are static equivalent, but this also implies that Σ is ruled from point (c) of Proposition 3.13 and the fact that the trivial system Σ' is ruled.

Static equivalence still appears explicitly in Theorem 4.2 because two general systems can be static equivalent without being ruled.

2. The part "$n > n'$ or $n < n'$" can be rephrased as follows: if a system is not ruled, it cannot be dynamic equivalent to any system of smaller dimension. No

necessary condition is given on the system of lower dimension; indeed, *any* system is dynamic equivalent to at least its first prolongation; see example 2 in section 3.5.

3. The case $n = n'$ states that dynamic equivalence, except when it reduces to static equivalence, forces both systems to be ruled (in the real analytic case, the added rigidity prevents the two situations from occurring simultaneously).

In other words, if two systems are not static equivalent and at least one of them is not ruled, they are not dynamic equivalent. Since the two conditions can be checked rather systematically, this yields a new and powerful method for proving that two systems are *not* dynamic equivalent, a difficult task in general because very few invariants of dynamic equivalence are known.

For instance, to the best of our knowledge, the state of the art does not allow one to decide whether Σ and Σ' in (23) are dynamic equivalent or not. In section 3.5, it was noted that they are not static equivalent and Σ' is not ruled. This implies the following corollary.

COROLLARY 4.3. Σ *and* Σ' *in* (23) *are not dynamic equivalent over any domains.*

4. Since being ruled is nongeneric [17], we have the following general consequence (in terms of germs of systems because the conclusion in the theorem is only local).

COROLLARY 4.4. *Generic static equivalence classes for germs of systems of the same dimension at a point are also dynamic equivalence classes.*

Note that this is in the mathematical sense of "generic": this does not prevent many interesting systems from being dynamic equivalent without being static equivalent. It might even be that the "most interesting systems" fall in this case!

5. Proofs.

Recall that subscripts always refer to the order of the jet space. The notation (3) is constantly used.

5.1. Preliminaries: A reformulation of dynamic and static equivalence.

The maps Φ and Ψ are always applied to jets of solutions, and, according to (12), the Kth jets of solutions of Σ remain in Σ_K; hence, the only information to retain about Φ and Ψ is their restriction to, respectively,

$$\widetilde{\Omega} = \Omega \cap \Sigma_K \quad \text{and} \quad \widetilde{\Omega}' = \Omega' \cap \Sigma'_{K'} . \tag{25}$$

We need one more piece of notation: according to section 2.3, the ℓth prolongation of a smooth map $\widetilde{\Phi} : \widetilde{\Omega} \to M'$ is a map $\pi_{K+\ell,\ell}^{-1}(\widetilde{\Omega}) \to J^\ell M'$; again, only its restriction to $\widetilde{\Omega}_{K+\ell}$ will matter; for this reason, the notations $\widetilde{\Phi}^{[\ell]}$ and $\widetilde{\Psi}^{[\ell]}$ will not stand for the prolongations as defined earlier but rather for these restrictions:

$$\widetilde{\Phi}^{[\ell]} : \widetilde{\Omega}_{K+\ell} \to J^\ell(M') , \quad \widetilde{\Psi}^{[\ell]} : \widetilde{\Omega'}_{K'+\ell} \to J^\ell(M) , \tag{26}$$

$$\text{with} \quad \widetilde{\Omega}_{K+\ell} = \Omega_{K+\ell} \cap \Sigma_{K+\ell} , \quad \widetilde{\Omega}'_{K'+\ell} = \Omega'_{K'+\ell} \cap \Sigma'_{K'+\ell} . \tag{27}$$

We may now state the following proposition. Smooth ($\mathbf{C}^\infty$ or $\mathbf{C}^\omega$) maps on $\widetilde{\Omega}_{K+\ell}$ or $\widetilde{\Omega'}_{K'+\ell}$ can be defined in a standard way because, from Proposition 3.3, these are smooth embedded submanifolds.

PROPOSITION 5.1 (dynamic equivalence). *Let* K, K' *be integers and* $\Omega \subset J^K(M)$ *and* $\Omega' \subset J^{K'}(M')$ *two open subsets. Systems* Σ *and* Σ' *are* dynamic equivalent over Ω and Ω' *if and only if, with* $\widetilde{\Omega}, \widetilde{\Omega}'$ *defined in* (25), *there exist two smooth (real analytic, in the real analytic case) mappings*

$$\widetilde{\Phi} : \widetilde{\Omega} \to M' \quad \textit{and} \quad \widetilde{\Psi} : \widetilde{\Omega}' \to M$$

such that

$$\widetilde{\Phi}^{[1]}\left(\widetilde{\Omega}_{K+1}\right) \subset \Sigma' \,, \quad \widetilde{\Psi}^{[1]}\left(\widetilde{\Omega'}_{K'+1}\right) \subset \Sigma \tag{28}$$

and, with $\widetilde{\Phi}^{[K]}$ *and* $\widetilde{\Psi}^{[K]}$ *defined by* (26),

$$\widetilde{\Phi}^{[K']}\left(\widetilde{\Omega}_{K+K'}\right) \subset \Omega' \,, \quad \widetilde{\Psi}^{[K]}\left(\widetilde{\Omega}'_{K+K'}\right) \subset \Omega \,, \tag{29}$$

$$\widetilde{\Psi} \circ \widetilde{\Phi}^{[K']} = \pi_{K+K',0}\Big|_{\widetilde{\Omega}_{K+K'}} \,, \quad \widetilde{\Phi} \circ \widetilde{\Psi}^{[K]} = \pi_{K+K',0}\Big|_{\widetilde{\Omega}'_{K+K'}} \,. \tag{30}$$

Proof. If the above conditions on Φ and Ψ are satisfied and $x(.) : I \to M$ is a solution of Σ whose Kth jet remains inside Ω, then the first part of (28) implies that $\mathcal{D}_{\Phi}^{K}(\,x(.)\,)$ is a solution of Σ', the first part of (29) implies that its Kth jet remains inside Ω', and the first part of (30) implies that $\mathcal{D}_{\Psi}^{K'}(\,\mathcal{D}_{\Phi}^{K}(\,x(.)\,)\,) = x(.)$. This proves the first item of Definition 3.8; the second item follows in the same way from the second part of (28), (29), and (30).

Conversely, if Φ and Ψ satisfy the properties of Definition 3.8, their restrictions $\widetilde{\Phi}$ and $\widetilde{\Psi}$ to $\widetilde{\Omega}$ and $\widetilde{\Omega'}$, respectively, satisfy the above relations because through each point in $\widetilde{\Omega}_{K+1}$, $\widetilde{\Omega'}_{K'+1}$, $\widetilde{\Omega}_{K+K'}$, or $\widetilde{\Omega}_{K+K'}$ passes a jet of order $K+1$, $K'+1$, or $K+K'$ of a solution of Σ or Σ'; differentiating yields the required relations. □

Proposition 5.2 (static equivalence). *With* $\Omega_1 \subset J^1(M) = \mathrm{T}M$ *and* $\Omega'_1 \subset J^1(M') = \mathrm{T}M$ *two open subsets, systems* Σ *and* Σ' *are* static equivalent over Ω_1 and Ω'_1 *if and only if, with* $\widetilde{\Omega}_1, \widetilde{\Omega}'_1$ *defined in* (25), *there exist a smooth diffeomorphism* $\Phi_0 : \widetilde{\Omega}_0 \to \widetilde{\Omega}'_0$ *and its inverse* Ψ_0 *such that* $\widetilde{\Phi}_0^{[1]}(\widetilde{\Omega}_1) = \widetilde{\Omega}'_1$ *(and* $\widetilde{\Psi}_0^{[1]}(\widetilde{\Omega}'_1) = \widetilde{\Omega}_1$*).*

Proof. The proof is a rephrasing of point (a) of Proposition 3.13. □

5.2. Proof of Theorem 4.2. Assume that Σ and Σ' are dynamic equivalent over the open sets $\Omega \subset J^K(M)$ and $\Omega' \subset J^{K'}(M')$; let $\widetilde{\Phi} : \widetilde{\Omega} \to M'$ and $\widetilde{\Psi} : \widetilde{\Omega}' \to M$ be the smooth maps given by Proposition 5.1 (recall that $\widetilde{\Omega}$ and $\widetilde{\Omega}'$ are open subsets of Σ_K and $\Sigma'_{K'}$). We define open subsets $\widetilde{\Omega}^S \subset \widetilde{\Omega}$ and $\widetilde{\Omega}'^S \subset \widetilde{\Omega}'$ and state four lemmas concerning these:

$$\xi \in \widetilde{\Omega}^S \quad \Leftrightarrow \quad \begin{array}{l} \text{There is a neighborhood } V \text{ of } \xi \text{ in } \widetilde{\Omega} \text{ and a smooth map} \\ \widetilde{\Phi}_0 : V_0 \to M' \text{ such that } \widetilde{\Phi}\Big|_V = \widetilde{\Phi}_0 \circ \pi_{K,0} \,, \end{array} \tag{31}$$

$$\xi' \in \widetilde{\Omega}'^S \quad \Leftrightarrow \quad \begin{array}{l} \text{There is a neighborhood } V' \text{ of } \xi' \text{ in } \widetilde{\Omega}' \text{ and a smooth map} \\ \widetilde{\Psi}_0 : V'_0 \to M \text{ such that } \widetilde{\Psi}\Big|_{V'} = \widetilde{\Psi}_0 \circ \pi_{K,0} \,. \end{array} \tag{32}$$

Lemma 5.3. *In the analytic case and if* $\widetilde{\Omega} = \Omega \cap \Sigma$ *and* $\widetilde{\Omega}' = \Omega' \cap \Sigma'$ *are connected, one has either* $\widetilde{\Omega}^S = \widetilde{\Omega}$ *or* $\widetilde{\Omega}^S = \varnothing$*, and either* $\widetilde{\Omega}'^S = \widetilde{\Omega}'$ *or* $\widetilde{\Omega}'^S = \varnothing$*.*

Lemma 5.4. *One has the following identities, where the two first ones hold for any subsets* $S \subset \widetilde{\Omega}$*,* $S' \subset \widetilde{\Omega}'$ *and any integer* ℓ*,* $0 \leq \ell \leq K + K'$*:*

$$\pi_{K+K',\ell}\left(\widetilde{\Phi}^{[K']^{-1}}(S')\right) = \widetilde{\Psi}^{[\ell]}\left(S'_{K'+\ell}\right), \quad \pi_{K+K',\ell}\left(\widetilde{\Psi}^{[K]^{-1}}(S)\right) = \widetilde{\Psi}^{[\ell]}\left(S_{K+\ell}\right), \tag{33}$$

$$\widetilde{\Phi}^{[1]}\left(\widetilde{\Omega}_{K+1}\right) = \widetilde{\Omega}'_1 \,, \qquad \widetilde{\Psi}^{[1]}\left(\widetilde{\Omega}'_{K'+1}\right) = \widetilde{\Omega}_1 \,. \tag{34}$$

Lemma 5.5. *If* $n < n'$*, then* $\widetilde{\Omega}^S = \varnothing$*. If* $n > n'$*, then* $\widetilde{\Omega}'^S = \varnothing$*.*

If $n = n'$, there are, for all $\xi_K \in \widetilde{\Omega}^S$, a neighborhood $\mathcal{V}_1$ of $\xi_1 = \pi_{K,1}(\xi_K)$ in Ω_1 and an open subset $\mathcal{V}'_1$ of Ω'_1 such that systems Σ and Σ' are static equivalent over $\mathcal{V}_1$ and $\mathcal{V}'_1$. There are also, for all $\xi'_{K'} \in \widetilde{\Omega}'^S$, a neighborhood $\mathcal{W}'$ of $\xi'_1 = \pi_{K',1}(\xi'_{K'})$ in Ω'_1 and an open subset $\mathcal{W}_1$ of Ω_1 such that systems Σ and Σ' are static equivalent over $\mathcal{W}_1$ and $\mathcal{W}'_1$. Finally,

$$\pi_{K+K',K'}\left(\widetilde{\Psi}^{[K]^{-1}}\left(\widetilde{\Omega}^S\right)\right) = \widetilde{\Phi}^{[K']}\left(\widetilde{\Omega}^S_{K+K'}\right) \;=\; \widetilde{\Omega}'^S\,, \tag{35}$$

$$\pi_{K+K',K}\left(\widetilde{\Phi}^{[K']^{-1}}\left(\widetilde{\Omega}'^S\right)\right) = \widetilde{\Psi}^{[K]}\left(\widetilde{\Omega}'^S_{K'+K}\right) \;=\; \widetilde{\Omega}^S\,. \tag{36}$$

LEMMA 5.6. *For all $\xi_{K+1} \in \widetilde{\Omega}_{K+1}$ such that $\xi_K = \pi_{K+1,K}(\xi_{K+1}) \in \widetilde{\Omega}\setminus\widetilde{\Omega}^S$, there is a straight line in $\mathrm{T}_{\widetilde{\Phi}(\xi_K)}M'$ that has contact of infinite order with Σ' at $\widetilde{\Phi}^{[1]}(\xi_{K+1})$.*

These lemmas will be proved later. Let us finish the proof of the theorem.

If $n < n'$, (34) implies the existence, for each $\xi' \in \widetilde{\Omega}'_1 = \Omega_1 \cap \Sigma'$, of some $\xi_{K+1} \in \widetilde{\Omega}_{K+1}$ such that $\widetilde{\Phi}^{[1]}(\xi_{K+1}) = \xi'$, and, finally, since $\widetilde{\Omega}^S$ is empty according to Lemma 5.5, Lemma 5.6 yields a straight line in $\mathrm{T}_{\xi'_0}M'$ that has contact of infinite order with Σ' at ξ'; from Proposition 3.7, this implies that system Σ' is ruled over Ω_1. If $n > n'$, one concludes in the same way.

Now assume $n = n'$. For all ξ' in $\widetilde{\Phi}^{[1]}((\widetilde{\Omega}\setminus\widetilde{\Omega}^S)_{K+1})$, there is, according to Lemma 5.6, a straight line in $\mathrm{T}_{\xi'_0}M'$ that has contact of infinite order with Σ' at ξ'. By continuity, this is also true for all ξ' in the topological closure

$$\widetilde{R}' \;=\; \overline{\widetilde{\Phi}^{[1]}\left(\left(\widetilde{\Omega}\setminus\widetilde{\Omega}^S\right)_{K+1}\right)} = \overline{\pi_{K+K',1}\left(\widetilde{\Psi}^{[K]^{-1}}\left(\widetilde{\Omega}\setminus\widetilde{\Omega}^S\right)\right)}\,, \tag{37}$$

where the second equality comes from (33). Let $i(\widetilde{R}')$ be the interior of $\widetilde{R}'$ for the induced topology on Σ'; since $\widetilde{R}' = \overline{i(\widetilde{R}')}$, there is an open subset $\mathcal{R}'$ of $\Omega'_1 \subset \mathrm{T}M'$, enjoying the property that it is the interior of its topological closure and such that $\mathcal{R}'\cap\Sigma' = i(\widetilde{R}')$ and $\overline{\mathcal{R}'}\cap\Sigma' = \widetilde{R}'$. From Proposition 3.7, Σ' is ruled over $\mathcal{R}'$. Setting $\mathcal{S}' = \Omega'_1 \setminus \overline{\mathcal{R}'}$, one has $\Omega'_1 = \overline{\mathcal{R}'}\cup\mathcal{S}' = \mathcal{R}'\cup\overline{\mathcal{S}'}$. Along the same lines, Σ is ruled over $\mathcal{R}$, open subset of $\Omega_1 \subset \mathrm{T}M$ such that $\mathcal{R}\cap\Sigma$ is the relative interior of

$$\widetilde{R} \;=\; \overline{\widetilde{\Psi}^{[1]}\left(\left(\widetilde{\Omega}'\setminus\widetilde{\Omega}'^S\right)_{K'+1}\right)} = \overline{\pi_{K+K',1}\left(\widetilde{\Phi}^{[K']^{-1}}\left(\widetilde{\Omega}'\setminus\widetilde{\Omega}'^S\right)\right)} \tag{38}$$

and such that $\Omega_1 = \overline{\mathcal{R}}\cup\mathcal{S} = \mathcal{R}\cup\overline{\mathcal{S}}$ with $\mathcal{S} = \Omega_1\setminus\mathcal{R}$.

We have proved (24) and point 1 of Theorem 4.2; let us prove point 2. Obviously,

$$\mathcal{S}\cap\Sigma \subset \pi_{K+K',1}\left(\widetilde{\Phi}^{[K']^{-1}}\left(\widetilde{\Omega}'^S\right)\right) \quad\text{and}\quad \mathcal{S}'\cap\Sigma' \subset \pi_{K+K',1}\left(\widetilde{\Psi}^{[K]^{-1}}\left(\widetilde{\Omega}^S\right)\right)\;.$$

Using identities (35) and (36), this implies

$$\mathcal{S}\cap\Sigma \subset \pi_{K,1}\left(\widetilde{\Omega}^S\right) \quad\text{and}\quad \mathcal{S}'\cap\Sigma' \subset \pi_{K',1}\left(\widetilde{\Omega}'^S\right)\;. \tag{39}$$

For all ξ in $\mathcal{S}\cap\Sigma$, there are one $\xi_K \in \widetilde{\Omega}^S$ such that $\xi = \pi_{K,1}(\xi_K)$ and, from Lemma 5.5, a neighborhood $\mathcal{V}_1^\xi$ of ξ in Ω_1 and an open subset $\mathcal{V}_1'^\xi$ of Ω'_1 such that systems Σ and Σ' are static equivalent over $\mathcal{V}_1^\xi$ and $\mathcal{V}_1'^\xi$. For all ξ' in $\mathcal{S}'\cap\Sigma'$, there are one $\xi'_{K'} \in \widetilde{\Omega}'^S$ such that $\xi' = \pi_{K',1}(\xi'_{K'})$ and, from Lemma 5.5, a neighborhood

$\mathcal{W}'^{\xi'}$ of $\xi'_1 = \pi_{K',1}(\xi'_{K'})$ in Ω'_1 and an open subset $\mathcal{W}_1^{\xi'}$ of Ω_1 such that systems Σ and Σ' are static equivalent over $\mathcal{W}_1^{\xi'}$ and $\mathcal{W}_1'^{\xi}$.

Now, $(\mathcal{V}_1^{\xi})_{\xi\in\mathcal{S}\cap\Sigma}$ is an open covering of $\mathcal{S}\cap\Sigma$ and $(\mathcal{W}_1'^{\xi'})_{\xi'\in\mathcal{S}'\cap\Sigma'}$ is an open covering of $\mathcal{S}'\cap\Sigma'$. Take for $(\widetilde{\mathcal{S}}^{\alpha})_{\alpha\in A}$ the union of $(\mathcal{V}_1^{\xi})_{\xi\in\mathcal{S}\cap\Sigma}$ and $(\mathcal{W}_1^{\xi'})_{\xi'\in\mathcal{S}'\cap\Sigma'}$; take for $(\widetilde{\mathcal{S}'}^{\alpha})_{\alpha\in A}$ the union of $(\mathcal{V}_1'^{\xi})_{\xi\in\mathcal{S}\cap\Sigma}$ and $(\mathcal{W}_1'^{\xi'})_{\xi'\in\mathcal{S}'\cap\Sigma'}$.

This proves the smooth case and obviously implies the real analytic one from Lemma 5.3.

Let us now prove the four lemmas used in the above proof.

Proof of Lemma 5.3. If $\widetilde{\Omega}^S \neq \varnothing$, then there is at least an open subset of $\widetilde{\Omega}$ in which the derivatives of $\widetilde{\Phi}$ along any vertical vector field (preserving fibers of $\Sigma_K \to M$) are identically zero; since these are real analytic, they must be zero all over $\widetilde{\Omega}$, assumed connected, and, hence, $\widetilde{\Omega}^S = \widetilde{\Omega}$. The proof is similar in $\widetilde{\Omega}'$. □

Proof of Lemma 5.4. The first relation in (33) is a consequence of the two identities

$$(40)\quad \pi_{K'+\ell,K'}\circ\widetilde{\Phi}^{[K'+\ell]} = \widetilde{\Phi}^{[K']}\circ\pi_{K+K'+\ell,K+K'} \quad\text{and}\quad \widetilde{\Psi}^{[\ell]}\circ\widetilde{\Phi}^{[K'+\ell]} = \pi_{K+K'+\ell,\ell}\,,$$

respectively, (6) with $(r,s) = (K'+\ell, K')$, and the ℓth prolongation of (30). The second relation follows from interchanging K, Φ, S with K', Ψ, S'.

From (28) and (29), one has, for any positive integer ℓ,

$$(41)\qquad \widetilde{\Phi}^{[\ell]}\left(\widetilde{\Omega}_{K+\ell}\right) \subset \widetilde{\Omega}'_\ell \quad\text{and}\quad \widetilde{\Psi}^{[\ell]}\left(\widetilde{\Omega}'_{K'+\ell}\right) \subset \widetilde{\Omega}_\ell$$

(for instance, (28) implies $\widetilde{\Phi}^{[\ell]}(\widetilde{\Omega}_{K+\ell}) \subset \Sigma'_\ell$, (29) implies $\widetilde{\Phi}^{[\ell]}(\widetilde{\Omega}_{K+\ell}) \subset \Omega'_\ell$, and, hence, the first relation above because $\widetilde{\Omega}'_\ell = \Omega'_\ell \cap \Sigma'_\ell$). We need only to prove the reverse inclusions for $\ell = 1$. Let us do it for the second one. The second relation in (40) for $\ell = 1$ implies $\widetilde{\Omega}_1 = \widetilde{\Psi}^{[1]}(\widetilde{\Phi}^{[K'+1]}(\widetilde{\Omega}_{K+K'+1}))$, and finally $\widetilde{\Omega}_1 \subset \widetilde{\Psi}^{[1]}(\widetilde{\Omega}'_{K'+1})$ from the first relation in (40) with $\ell = K'+1$. □

Proof of Lemma 5.5. Assume, for instance, that $\widetilde{\Omega}^S$ is nonempty; then it contains an open subset V and there is a smooth $\widetilde{\Phi}_0 : V_0 \to M'$ such that, in restriction to V, $\widetilde{\Phi} = \widetilde{\Phi}_0 \circ \pi_{K,0}$. Hence, (30) implies, on the open subset $V' = (\widetilde{\Psi}^{[K]})^{-1}(V)$ of $\Sigma'{}_{K+K'}$,

$$(42)\qquad \widetilde{\Phi}_0 \circ \pi_{K,0} \circ \widetilde{\Psi}^{[K]} = \pi_{K+K',0}\big|_{V'}\;.$$

The rank of the map on the left-hand side is n' while the rank on the right-hand side is no larger than n (rank of $\pi_{K,0}$), and, hence, $\widetilde{\Omega}^S \neq \varnothing$ implies $n' \leq n$. By interchanging the two systems, this proves the fist sentence of the lemma.

Let us now turn to the case where $n = n'$. Consider ξ_K in $\widetilde{\Omega}^S$. By definition of $\widetilde{\Omega}^S$, there are a neighborhood V and a smooth (real analytic in the real analytic case) map $\widetilde{\Phi}_0 : V_0 \to M'$ such that $\widetilde{\Phi} = \widetilde{\Phi}_0 \circ \pi_{K,0}$ on V. Let V' be defined from V as

$$(43)\qquad V' = \pi_{K+K',K'}\left({\widetilde{\Psi}^{[K]}}^{-1}(V)\right) = \widetilde{\Phi}^{[K']}(V_{K+K'})\,,$$

where the second equality comes from (33). Applying $\widetilde{\Psi}$ and $\widetilde{\Psi}^{[1]}$ to both sides of the first equality in (6) and using (43) with $(r,s) = (K,0)$ and $(r,s) = (K,1)$ yields

$$(44)\qquad \widetilde{\Psi}(V') = V_0\,,\qquad \widetilde{\Psi}^{[1]}(V'_{K'+1}) = V_1\,.$$

Substituting $\widetilde{\Phi} = \widetilde{\Phi}_0 \circ \pi_{K,0}$ in (30), one has $\widetilde{\Phi}_0 \circ \widetilde{\Psi} \circ \pi_{K+K',K'} = \pi_{K+K',0}$ on ${\widetilde{\Psi}^{[K]}}^{-1}(V)$, and, finally,

$$(45)\qquad \widetilde{\Phi}_0 \circ \widetilde{\Psi} = \pi_{K',0} \text{ on } V'\,;$$

in a similar way, substituting $\widetilde{\Phi}^{[1]} = \widetilde{\Phi}_0^{[1]} \circ \pi_{K+1,1}$ in the first prolongation of (30),

$$\widetilde{\Phi}_0^{[1]} \circ \widetilde{\Psi}^{[1]} = \pi_{K'+1,1} \text{ on } V'_{K'+1}\,. \tag{46}$$

Applying $\widetilde{\Phi}_0$ to both sides of the first relation and $\widetilde{\Phi}_0^{[1]}$ to both sides of the second relation in (44), one has, using (45) and (46),

$$\widetilde{\Phi}_0(V_0) = V'_0\,, \quad \widetilde{\Phi}_0^{[1]}(V_1) = V'_1\,. \tag{47}$$

Since the rank of $\pi_{K',0}$ on the right-hand side of (45) is $n' = n$ at all points of V', $\widetilde{\Phi}_0$ must be a local diffeomorphism at all points of $\widetilde{\Psi}(V') = V_0$ and, in particular, at ξ_0: by the inverse function theorem, there are a neighborhood O of $\xi_0 = \pi_{K,0}(\xi)$ in V_0 and a neighborhood O' of $\Phi_0(\xi_0)$ in M' such that Φ_0 defines a diffeomorphism $O \to O'$.

Let us now replace V with $V \cap {\pi_{K,0}}^{-1}(O)$, a smaller neighborhood of ξ_K; V' is still defined by (43) from this smaller V, one has $V_0 = O$, the former $\widetilde{\Phi}_0$ is replaced by its restriction to this smaller V_0, and the above relations still hold. In particular, $O' = \widetilde{\Phi}_0(O)$ must be all V'_0 according to (47); i.e., $\widetilde{\Phi}_0$ defines a diffeomorphism $V_0 \to V'_0$; let $\widetilde{\Psi}_0$ be its inverse. Composing each side of (45) with $\widetilde{\Psi}_0$, one gets $\widetilde{\Psi} = \widetilde{\Psi}_0 \circ \pi_{K',0}$ on V'; hence, by (32), one has $V' \subset \widetilde{\Omega}'^S$ and, since this is true for all ξ_K in $\widetilde{\Omega}^S$, one has

$$\pi_{K+K',K'}\left(\widetilde{\Psi}^{[K]^{-1}}(\widetilde{\Omega}^S)\right) = \widetilde{\Phi}^{[K']}(\widetilde{\Omega}^S_{K+K'}) \subset \widetilde{\Omega}'^S\,. \tag{48}$$

Let $\mathcal{V}_1$ and $\mathcal{V}'_1$ be open subsets of Ω_1 and Ω'_1 such that

$$V_1 = \Sigma \cap \mathcal{V}_1\,, \quad V'_1 = \Sigma \cap \mathcal{V}'_1\,. \tag{49}$$

From Proposition 5.2, the second relation in (47) implies that systems Σ and Σ' are static equivalent over $\mathcal{V}_1$ and $\mathcal{V}'_1$. Interchanging the two systems, one proves that

$$\pi_{K+K',K}\left(\widetilde{\Phi}^{[K']^{-1}}(\widetilde{\Omega}'^S)\right) = \widetilde{\Psi}^{[K]}(\widetilde{\Omega}'^S_{K+K'}) \subset \widetilde{\Omega}^S \tag{50}$$

and that, for all $\xi'_{K'} \in \widetilde{\Omega}'^S$, there are a neighborhood $\mathcal{W}'$ of $\xi'_1 = \pi_{K',1}(\xi'_{K'})$ in Ω'_1 and an open subset $\mathcal{W}_1$ of Ω_1 such that systems Σ and Σ' are static equivalent over $\mathcal{W}_1$ and $\mathcal{W}'_1$.

Now, $\widetilde{\Phi}^{[K']}(\widetilde{\Omega}^S_{K+K'}) \subset \widetilde{\Omega}'^S$ in (48) implies $\widetilde{\Omega}^S_{K+K'} \subset \widetilde{\Phi}^{[K']^{-1}}(\widetilde{\Omega}'^S)$, and, hence, $\widetilde{\Omega}^S \subset \pi_{K+K',K}(\widetilde{\Phi}^{[K']^{-1}}(\widetilde{\Omega}'^S))$. Hence, (48) implies the converse inclusion in (50); in a similar way (50) implies the converse inclusion in (48). This proves (35) and (36) and ends the proof of Lemma 5.5. □

Proof of Lemma 5.6. Denote by $\bar{\xi}_{K+1}$ the point ξ_{K+1} in the lemma statement, and set $\bar{\xi}_K = \pi_{K+1,K}(\bar{\xi}_{K+1}) \in \widetilde{\Omega} \setminus \widetilde{\Omega}^S$, $\bar{\xi}_0 = \pi_{K,0}(\bar{\xi}_{K+1})$, and $\bar{\xi}_1 = \pi_{K,1}(\bar{\xi}_{K+1})$. From Proposition 3.4 and after possibly shrinking $\mathcal{U}_K$ so that it is contained in Ω, there exist a neighborhood $\mathcal{U}_K \subset \Omega$ of $\bar{\xi}_K$ in $J^K(M)$, coordinates $(x_{\mathrm{I}}, x_{\mathbb{I}})$ on $\mathcal{U}_0 = \pi_{K,0}(\mathcal{U}_K)$ inducing coordinates $(x_{\mathrm{I}}, x_{\mathbb{I}}, \dot{x}_{\mathrm{I}}, \dot{x}_{\mathbb{I}}, \ldots, x_{\mathrm{I}}^{(K)}, x_{\mathbb{I}}^{(K)})$ on $\mathcal{U}_K$, and an open subset $U_K \subset \mathbb{R}^{n+Km}$ such that the equations of $\widetilde{\mathcal{U}}_K = \mathcal{U}_K \cap \Sigma_K$ in $J^K(M)$ in these coordinates are

$$\begin{aligned} x_{\mathrm{I}}^{(i)} &= f^{(i-1)}(x_{\mathrm{I}}, x_{\mathbb{I}}, \dot{x}_{\mathbb{I}}, \ldots, x_{\mathbb{I}}^{(i)})\,, \quad 1 \le i \le K\,, \\ (x_{\mathrm{I}}, x_{\mathbb{I}}, \dot{x}_{\mathbb{I}}, \ldots, x_{\mathbb{I}}^{(K)}) &\in U_K\,. \end{aligned} \tag{51}$$

By substitution, there is a unique smooth map $\phi_K : U_K \to M'$ such that $\widetilde{\Phi}(\xi) = \phi_K(x_{\mathrm{I}}, x_{\mathbb{I}}, \dot{x}_{\mathbb{I}}, \dots, x_{\mathbb{I}}^{(K)})$ for all ξ in $\widetilde{\mathcal{U}}_K$ with coordinate vector $(x_{\mathrm{I}}, x_{\mathbb{I}}, \dots, x_{\mathrm{I}}^{(K)}, x_{\mathbb{I}}^{(K)})$.

Let $\overline{X}_i = (\overline{x_{\mathrm{I}}}, \overline{x_{\mathbb{I}}}, \overline{\dot{x}_{\mathrm{I}}}, \overline{\dot{x}_{\mathbb{I}}}, \dots, \overline{x_{\mathrm{I}}^{(i)}}, \overline{x_{\mathbb{I}}^{(i)}})$ be the coordinate vector of $\bar{\xi}_i$ for $i \leq K+1$ and $\bar{\rho}$ the smallest integer such that ϕ_K does not depend on $x_{\mathbb{I}}^{(\bar{\rho}+1)}, \dots, x_{\mathbb{I}}^{(K)}$ on at least one neighborhood of $\overline{X}_K$. Shrinking U_K to this neighborhood and $\widetilde{\mathcal{U}}_K$ accordingly, we may define $\phi : U_{\bar{\rho}} \to M'$, with $U_{\bar{\rho}}$ the projection of U_K on $\mathbb{R}^{n+\bar{\rho}\,m}$, such that $\widetilde{\Phi}(\xi) = \phi_K(x_{\mathrm{I}}, x_{\mathbb{I}}, \dot{x}_{\mathbb{I}}, \dots, x_{\mathbb{I}}^{(K)}) = \phi(x_{\mathrm{I}}, x_{\mathbb{I}}, \dot{x}_{\mathbb{I}}, \dots, x_{\mathbb{I}}^{(\bar{\rho})})$. If $\bar{\rho}$ was zero, one would have $\widetilde{\Phi}(\xi) = \phi(x_{\mathrm{I}}, x_{\mathbb{I}})$, and, hence, the right-hand side of (31) would be satisfied for $\xi = \bar{\xi}_K$ with $V = \widetilde{\mathcal{U}}_K$; this is impossible because we assumed $\bar{\xi}_K \in \widetilde{\Omega} \setminus \widetilde{\Omega}^S$. Hence, $\bar{\rho} \geq 1$.

For all ξ_{K+1} in $\widetilde{\mathcal{U}}_{K+1}$ with coordinate vector $(x_{\mathrm{I}}, x_{\mathbb{I}}, \dots, x_{\mathbb{I}}^{(K+1)})$, one has

$$\widetilde{\Phi}^{[1]}(\xi_{K+1}) = \chi\left(x_{\mathrm{I}}, x_{\mathbb{I}}, \dot{x}_{\mathbb{I}}, \dots, x_{\mathbb{I}}^{(\bar{\rho})}, x_{\mathbb{I}}^{(\bar{\rho}+1)}\right) \tag{52}$$

with $\chi : U_{\bar{\rho}+1} \to \mathrm{T}M'$ the map defined by

(53)

$$\chi\left(x_{\mathrm{I}}, \dots, x_{\mathbb{I}}^{(\bar{\rho}+1)}\right) = \left(\phi\left(x_{\mathrm{I}}, \dots, x_{\mathbb{I}}^{(\bar{\rho})}\right),\ a\left(x_{\mathrm{I}}, \dots, x_{\mathbb{I}}^{(\bar{\rho})}\right) + \frac{\partial \phi}{\partial x_{\mathbb{I}}^{(\bar{\rho})}}\left(x_{\mathrm{I}}, \dots, x_{\mathbb{I}}^{(\bar{\rho})}\right) x_{\mathbb{I}}^{(\bar{\rho}+1)} \right)$$

with $a = \frac{\partial \phi}{\partial x_{\mathrm{I}}} f + \sum_{i=0}^{\bar{\rho}-1} \frac{\partial \phi}{\partial x_{\mathbb{I}}^{(i)}} x_{\mathbb{I}}^{(i+1)}$. According to (29), (51), and (52), Σ' contains $\chi(U_{\bar{\rho}+1})$. Now, for any $(x_{\mathrm{I}}, \dots, x_{\mathbb{I}}^{(\bar{\rho}+1)}) \in U_{\bar{\rho}+1}$ such that the linear map

$$\frac{\partial \phi}{\partial x_{\mathbb{I}}^{(\rho)}}\left(x_{\mathrm{I}}, \dots, x_{\mathbb{I}}^{(\bar{\rho})}\right) : \quad \mathbb{R}^m \ \to\ \mathrm{T}_{\phi\left(x_{\mathrm{I}}, \dots, x_{\mathbb{I}}^{(\bar{\rho})}\right)} M'$$

is nonzero, picking $\underline{w} \neq 0$ in its range, (53) implies that the straight line Δ in $\mathrm{T}_{\phi(x_{\mathrm{I}}, \dots, x_{\mathbb{I}}^{(\bar{\rho})})} M'$ passing through $\chi(x_{\mathrm{I}}, \dots, x_{\mathbb{I}}^{(\bar{\rho}+1)})$ with direction $\underline{w}$ has a segment around $\chi(x_{\mathrm{I}}, \dots, x_{\mathbb{I}}^{(\bar{\rho}+1)})$ contained in Σ', and, hence, in particular, Δ has contact of infinite order with Σ' at point $\chi(x_{\mathrm{I}}, \dots, x_{\mathbb{I}}^{(\bar{\rho}+1)})$. To sum up, we have proved so far that, for all ξ_{K+1} in $\widetilde{\mathcal{U}}_{K+1}$ with coordinate vector $(x_{\mathrm{I}}, x_{\mathbb{I}}, \dots, x_{\mathbb{I}}^{(K+1)})$ such that $\frac{\partial \phi}{\partial x_{\mathbb{I}}^{(\rho)}}(x_{\mathrm{I}}, \dots, x_{\mathbb{I}}^{(\bar{\rho})})$ is nonzero, there is a straight line $\Delta_{\xi_{K+1}}$ in $\mathrm{T}_{\widetilde{\Phi}(\xi_K)} M'$ passing through $\widetilde{\Phi}^{[1]}(\xi_{K+1})$ that has contact of infinite order with Σ' at $\widetilde{\Phi}^{[1]}(\xi_{K+1})$. The set of such points ξ_{K+1} may not contain $\bar{\xi}_{K+1}$, but its topological closure does by minimality of $\bar{\rho}$; taking a sequence of points ξ_{K+1} that converges to $\bar{\xi}_{K+1}$, any accumulation point of the compact sequence $(\Delta_{\xi_{K+1}})$ is a straight line in $\mathrm{T}_{\widetilde{\Phi}(\bar{\xi}_K)} M'$ passing through $\widetilde{\Phi}^{[1]}(\bar{\xi}_{K+1})$ that has contact of infinite order with Σ' at $\widetilde{\Phi}^{[1]}(\bar{\xi}_{K+1})$. □

Acknowledgment. The author is indebted to the anonymous reviewer who suggested to include a sketch of the proof (section 4.2) that clarifies the main ideas.

REFERENCES

[1] R. L. Anderson and N. H. Ibragimov, *Lie-Bäcklund Transformations in Applications*, SIAM Stud. Appl. Math. 1, SIAM, Philadelphia, 1979.

[2] D. Avanessoff, *Linéarisation Dynamique des Systèmes non Linéaires et Paramétrage de l'Ensemble des Solutions*, Ph.D. thesis, University de Nice - Sophia Antipolis, Nice, France, 2005.

[3] D. Avanessoff and J.-B. Pomet, *Flatness and Monge parameterization of two-input systems, control-affine with* 4 *states or general with* 3 *states*, ESAIM Control Optim. Calc. Var., 13 (2007), pp. 237–264.

[4] É. Cartan, *Sur l'équivalence absolue de certains systèmes d'équations différentielles et sur certaines familles de courbes*, Bull. Soc. Math. France, 42 (1914), pp. 12–48.

[5] B. Charlet, J. Lévine, and R. Marino, *Sufficient conditions for dynamic state feedback linearization*, SIAM J. Control Optim., 29 (1991), pp. 38–57.

[6] M. Fliess, J. Lévine, P. Martin, and P. Rouchon, *Sur les systèmes non linéaires différentiellement plats*, C. R. Acad. Sci. Paris Sér. I, 315 (1992), pp. 619–624.

[7] M. Fliess, J. Lévine, P. Martin, and P. Rouchon, *A Lie-Bäcklund approach to equivalence and flatness of nonlinear systems*, IEEE Trans. Automat. Control, 44 (1999), pp. 922–937.

[8] M. Golubitsky and V. Guillemin, *Stable Mappings and Their Singularities*, Grad. Texts in Math. 14, Springer-Verlag, New York, 1973.

[9] E. Goursat, *Sur le problème de Monge*, Bull. Soc. Math. France, 33 (1905), pp. 201–210.

[10] D. Hilbert, *Über den begriff der klasse von differentialgleichungen*, Math. Ann., 73 (1912), pp. 95–108.

[11] A. Isidori, C. H. Moog, and A. de Luca, *A sufficient condition for full linearization via dynamic state feedback*, in Proceedings of the 25th IEEE Conference on Decision & Control, Athens, 1986, pp. 203–207.

[12] B. Jakubczyk, *Equivalence of differential equations and differential algebras*, Tatra Mt. Math. Publ., 4 (1994), pp. 125–130.

[13] I. Kupka, *On feedback equivalence*, in Proceedings of the Canadian Mathematical Society Conference Differential Geometry, Global Analysis, and Topology Halifax, NS, Canada, (1990), Amer. Math. Soc., 1991, pp. 105–117.

[14] J. M. Landsberg, *Is a linear space contained in a submanifold? On the number of derivatives needed to tell*, J. Reine Angew. Math., 508 (1999), pp. 53–60.

[15] P. Martin, *Contribution à l'Étude des Systèmes Differentiellement Plats*, Ph.D. thesis, L'École Nationale Supérieure des Mines de Paris, Paris, 1992.

[16] J.-B. Pomet, *A differential geometric setting for dynamic equivalence and dynamic linearization*, in Geometry in Nonlinear Control and Differential Inclusions, Banach Center Publ. 32, Polish Academy of Sciences, Warsaw, 1995, pp. 319–339.

[17] P. Rouchon, *Necessary condition and genericity of dynamic feedback linearization*, J. Math. Syst. Estimation Control, 4 (1994), pp. 1–14.

[18] W. Shadwick, *Absolute equivalence and dynamic feedback linearization*, Systems Control Lett., 15 (1990), pp. 35–39.

[19] W. F. Shadwick and W. M. Sluis, *Dynamic feedback for classical geometries*, in Differential Geometry and Mathematical Physics (Vancouver, BC, 1993), Contemp. Math. 170, Amer. Math. Soc., Providence, RI, 1994, pp. 207–213.

[20] W. M. Sluis, *A necessary condition for dynamic feedback linearization*, Systems & Control Lett., 21 (1993), pp. 277–283.

[21] K. Tchoń, *The only stable normal forms of affine systems under feedback are linear*, Systems & Control Lett., 8 (1987), pp. 359–365.

[22] G. R. Wilkens, *Centro-affine geometry in the plane and feedback invariants of two-state scalar control systems*, in Differential Geometry and Control (Boulder, CO, 1997), Proc. Sympos. Pure Math. 64, Amer. Math. Soc., Providence, RI, 1999, pp. 319–333.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 2, pp. 941–971

© 2009 Society for Industrial and Applied Mathematics

OPTIMAL STOPPING PROBLEM FOR STOCHASTIC DIFFERENTIAL EQUATIONS WITH RANDOM COEFFICIENTS[*]

MOU-HSIUNG CHANG[†], TAO PANG[‡], AND JIONGMIN YONG[§]

Abstract. An optimal stopping problem for stochastic differential equations with random coefficients is considered. The dynamic programming principle leads to a Hamilton–Jacobi–Bellman equation, which, for the current case, is a backward stochastic partial differential variational inequality (BSPDVI, for short) for the value function. Well-posedness of such a BSPDVI is established, and a verification theorem is proved.

Key words. optimal stopping, random coefficients, dynamic programming principle, backward stochastic partial differential variational inequality, verification theorem

AMS subject classifications. 49L20, 93E20

DOI. 10.1137/070705726

1. Introduction.

Throughout this paper, we let $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ be a complete filtered probability space on which a d-dimensional standard Brownian motion $W(\cdot)$ is defined, with $\mathbb{F} \equiv \{\mathcal{F}_t\}_{t\ge 0}$ being its natural filtration augmented by all the $\mathbb{P}$-null sets in $\mathcal{F}$. Let $\mathcal{S}[0,T]$ be the set of all $\mathbb{F}$-stopping times taking values in $[0,T]$. For any $\tau_1, \tau_2 \in \mathcal{S}[0,T]$, with $\tau_1 \le \tau_2$ almost surely and $\mathbb{P}\{\tau_1 < \tau_2\} > 0$, let

$$
(1.1)\qquad \begin{cases} \mathcal{S}[\tau_1,\tau_2] \overset{\Delta}{=} \{\tau \in \mathcal{S}[0,T] | \tau_1 \le \tau \le \tau_2 \text{ a.s. }\}, \\ \mathcal{S}(\tau_1,\tau_2] \overset{\Delta}{=} \{\tau \in \mathcal{S}[\tau_1,\tau_2] | \tau_1 < \tau \text{ a.s. } \text{ on } \{\tau_1 < \tau_2\}\}, \\ \mathcal{S}[\tau_1,\tau_2) \overset{\Delta}{=} \{\tau \in \mathcal{S}[\tau_1,\tau_2] | \tau < \tau_2 \text{ a.s. } \text{ on } \{\tau_1 < \tau_2\}\}, \\ \mathcal{S}(\tau_1,\tau_2) \overset{\Delta}{=} \{\tau \in \mathcal{S}[\tau_1,\tau_2] | \tau_1 < \tau < \tau_2 \text{ a.s. } \text{ on } \{\tau_1 < \tau_2\}\}. \end{cases}
$$

Next, for any $s \in \mathcal{S}[0,T]$ and $p \ge 1$, denote

$$
(1.2)\qquad \mathcal{X}_s^p \equiv L^p_{\mathcal{F}_s}(\Omega;\mathbb{R}^n) \overset{\Delta}{=} \left\{\xi : \Omega \to \mathbb{R}^n | \xi \text{ is } \mathcal{F}_s\text{-measurable}, \mathbb{E}|\xi|^p < \infty\right\}.
$$

For any $s \in \mathcal{S}[0,T)$ and $\xi \in \mathcal{X}_s^p$, consider the following stochastic differential equation (SDE, for short):

$$
(1.3)\qquad \begin{cases} dX(t) = b(t,X(t))dt + \sigma(t,X(t))dW(t), \qquad t \in [s,T], \\ X(s) = \xi, \end{cases}
$$

where $b : [0,T] \times \mathbb{R}^n \times \Omega \to \mathbb{R}^n$ and $\sigma : [0,T] \times \mathbb{R}^n \times \Omega \to \mathbb{R}^{n\times d}$ are given maps. We refer to the above as the *state equation.* Under proper conditions (which will be

[*] Received by the editors October 18, 2007; accepted for publication (in revised form) November 10, 2008; published electronically February 25, 2009. This work is supported in part by grant W911NF-04-D-0003 from the U.S. Army Research Office and by NSF grant DMS-0604309.
http://www.siam.org/journals/sicon/48-2/70572.html

[†] Mathematics Division, U.S. Army Research Office, Research Triangle Park, NC 27709 (mouhsiung.chang@us.army.mil).

[‡] Department of Mathematics, Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC 27695 (tpang@unity.ncsu.edu).

[§] Department of Mathematics, University of Central Florida, Orlando, FL 32816 (jyong@mail.ucf.edu).

assumed shortly), the above SDE admits a unique strong solution $X(\cdot) \equiv X(\cdot\,; s, \xi)$. Introduce the following *cost functional*:

$$J_{s,\xi}(\tau) = \mathbb{E}\left[\int_s^\tau g(t, X(t; s, \xi))dt + h(\tau, X(\tau; s, \xi))\big|\mathcal{F}_s\right], \qquad \tau \in \mathcal{S}[s, T], \tag{1.4}$$

where $g, h : [0, T]\times\mathbb{R}^n\times\Omega \to [0, \infty)$ are some given nonnegative maps satisfying proper conditions. The two terms on the right-hand side of (1.4) represent the *running cost* and the *terminal cost*, respectively. We point out that all the involved maps b, σ, g, and h in our discussion are allowed to be random. With the above setting, we can now pose the following optimal stopping problem.

PROBLEM (S). *For given $s \in \mathcal{S}[0, T)$ and $\xi \in \mathcal{X}_s^p$, find the smallest $\bar\tau \in \mathcal{S}[s, T]$ such that*

$$J_{s,\xi}(\bar\tau) = \inf_{\tau\in\mathcal{S}[s,T]} J_{s,\xi}(\tau) \equiv V(s, \xi). \tag{1.5}$$

Any $\bar\tau \in \mathcal{S}[s, T]$ satisfying (1.5) is referred to as an *optimal stopping time*, and the smallest one is referred to as the *smallest optimal stopping time.* We compatibly define

$$V(T, \xi) = h(T, \xi) \qquad \forall \xi \in \mathcal{X}_s^p. \tag{1.6}$$

Random field $V(\cdot\,, \cdot)$ defined by (1.5)–(1.6) is called the *value function* of Problem (S). We point out that for the maps g and h, a nonnegativity condition can be relaxed to the boundedness from below. On the other hand, it is not hard to see that if $h = 0$ and $g > 0$, then any optimal stopping time of Problem (S) must be the smallest one. But, in general, the optimal stopping time of Problem (S) is not necessarily unique (one can modify Example D.14 of [14]). Hence, to be definite, our Problem (S) is to find the smallest optimal stopping time. We also note that, due to the fact that the coefficients are allowed to be random and our cost functional is defined by a conditional expectation, our value function $V(\cdot\,, \cdot)$ is actually a random field.

In the case where all the coefficients are deterministic, one can prove the dynamic programming principle which leads to a partial differential variational inequality (PDVI, for short), as the corresponding HJB equation for the value function (which is deterministic). Moreover, it can be shown that the value function is the unique viscosity solution to the PDVI. In the case where the diffusion is uniformly nondegenerate, the value function is the (unique) classical solution of the PDVI, provided that some mild smoothness conditions are assumed for the coefficients. On the other hand, one can independently establish the well-posedness of the corresponding PDVI, as well as a verification theorem. These will then provide a solution to the original optimal stopping time problem (see [2] and the references cited therein).

We also note that, by some pure probabilistic approach, one can study an optimal stopping time problem for general continuous-time stochastic processes. Optimal stopping time is characterized by means of the so-called Snell's envelope, super martingale, and so on, without using the dynamic programming principle. In such an approach, no HJB equation is involved, which is natural because no dynamic equation is assumed for the considered stochastic processes (see [14]). We refer to [9], [10], [25], [3], [24], [8], [28], [22], [29], [18], [1], [5], [23], [4], [27], [7], [11] for relevant results on stochastic optimal stopping and optimal control problems.

For the problem under our consideration, since we have more structures on the stochastic process (satisfying an SDE, etc.), it is expected to have more detailed characterization on the optimal stopping time. On the other hand, due to the randomness of the coefficients, the usual technique of the dynamic programming principle together with theories of PDVIs, do not directly apply. In this paper, inspired by [21], we will formally derive the corresponding HJB equation for the value function $V(\cdot\,,\cdot)$, which is now a *backward stochastic partial differential variational inequality* (BSPDVI, for short). Using a result of semilinear backward stochastic partial differential equations (BSPDEs, for short) from [26], together with a standard penalty technique for (deterministic) PDVIs (see [12]), we will obtain the well-posedness of our BSPDVI in a certain sense. At the same time, a verification theorem will be established, which says that, under proper conditions, the solution to the BSPDVI coincides with the value function of Problem (S). Then an optimal stopping time can be characterized. See [20] for some results concerning backward stochastic variational inequalities in an abstract framework.

The rest of the paper is organized as follows: Some preliminary results, including certain basic properties of the value function will be presented in section 2. In section 3, we will formally derive the BSPDVI and formally prove a verification theorem. Notions of adapted solutions will be introduced in section 4. The well-posedness of the BSPDVI will be established in section 5. Finally, in section 6, the adapted weak solution of the BSPDVI will be identified as the value function of Problem (S).

2. Some preliminary results. In this section, we are going to present some preliminary results related to value function $V(\cdot\,,\cdot)$ of Problem (S). To begin with, for any $p \geq 1$, $s \in \mathcal{S}[0,T)$, and $\tau \in \mathcal{S}(s,T]$, we let $L^p_{\mathbb{F}}(\Omega; C([s,\tau];\mathbb{R}^n))$ be the set of all processes $\varphi : [s,\tau] \to \mathbb{R}^n$ having continuous paths and

$$\mathbb{E}\left[\sup_{t\in[s,\tau]} |\varphi(t)|^p\right] < \infty.$$

It is clear that $L^p_{\mathbb{F}}(\Omega; C([s,\tau];\mathbb{R}^n))$ is a Banach space. Next, for $p \geq 1$, we denote (recall (1.2))

$$\mathcal{D}^p = \left\{(s,\xi) \in \mathcal{S}[0,T] \times \mathcal{X}^p_T \;\middle|\; s \in \mathcal{S}[0,T],\ \xi \in \mathcal{X}^p_s\right\}.$$

Now, we introduce the following standing assumption concerning the coefficients of state equation (1.3).

Assumption (H1). Maps $b : [0,T] \times \mathbb{R}^n \times \Omega \to \mathbb{R}^n$ and $\sigma : [0,T] \times \mathbb{R}^n \times \Omega \to \mathbb{R}^{n\times d}$ are measurable and they satisfy the following:

(a) For each $x \in \mathbb{R}^n$, $t \mapsto (b(t,x), \sigma(t,x))$ is $\mathbb{F}$-progressively measurable and for some $p > 1$,

$$\mathbb{E}\left(\int_0^T |b(t,0)|dt\right)^p + \mathbb{E}\left(\int_0^T |\sigma(t,0)|^2 dt\right)^{\frac{p}{2}} < \infty. \tag{2.1}$$

(b) There exists an $L > 0$ such that

$$\begin{aligned} &|b(t,x,\omega) - b(t,y,\omega)| + |\sigma(t,x,\omega) - \sigma(t,y,\omega)| \leq L|x-y|, \\ &\qquad\qquad\qquad\qquad \text{a.e. } t \in [0,T],\ \forall x,y \in \mathbb{R}^n,\ \text{a.s. } \omega \in \Omega. \end{aligned} \tag{2.2}$$

Concerning the maps appearing in the cost functional, we introduce the following assumption.

Assumption (H2). Maps $g, h : [0,T] \times \mathbb{R}^n \times \Omega \to [0,\infty)$ are measurable, and they satisfy the following:

(a) For each $x \in \mathbb{R}^n$, $t \mapsto (g(t,x), h(t,x))$ is $\mathbb{F}$-progressively measurable; for each $x \in \mathbb{R}^n$ and a.s. $\omega \in \Omega$, $t \mapsto h(t,x)$ is continuous, and

$$\mathbb{E}\left[\int_0^T g(t,0)dt + \sup_{t\in[0,T]} h(t,0)\right] < \infty. \tag{2.3}$$

(b) There exists an $L > 0$ such that

$$\begin{aligned} &|g(t,x,\omega) - g(t,y,\omega)| + |h(t,x,\omega) - h(t,y,\omega)| \le L|x-y|, \\ &\qquad\qquad\qquad\qquad\qquad \text{a.e. } t \in [0,T],\ \forall x, y \in \mathbb{R}^n,\ \text{a.s. } \omega \in \Omega, \end{aligned} \tag{2.4}$$

and there exists a continuous nondecreasing function $\rho : [0,\infty) \to [0,\infty)$, with $\rho(0) = 0$ such that

$$|h(t,x,\omega) - h(s,x,\omega)| \le (1+|x|)\rho(|t-s|), \qquad \forall t, s \in [0,T],\ x \in \mathbb{R}^n, \quad \text{a.s.} \tag{2.5}$$

The following result is pretty standard (see [13]).

Proposition 2.1. *Let* (H1) *hold. Then, for each* $(s,\xi) \in \mathcal{D}^p$, *state equation* (1.3) *admits a unique (strong) solution* $X(\cdot) \equiv X(\cdot\,; s,\xi) \in L^p_{\mathbb{F}}(\Omega; C([s,T];\mathbb{R}^n))$. *Moreover,*

$$\mathbb{E}\left[\sup_{t\in[s,T]} |X(t;s,\xi)|^p \;\Big|\; \mathcal{F}_s\right] \le C\left(1+|\xi|^p\right) \qquad \forall (s,\xi) \in \mathcal{D}^p, \tag{2.6}$$

$$\mathbb{E}\left[\sup_{t\in[s,T]} |X(t;s,\xi) - X\left(t;s,\bar\xi\right)|^p \;\Big|\; \mathcal{F}_s\right] \le C\left|\xi - \bar\xi\right|^p \qquad \forall (s,\xi), \left(s,\bar\xi\right) \in \mathcal{D}^p, \tag{2.7}$$

and when $p > 1$,

$$\begin{aligned} &\mathbb{E}\left[\sup_{t\in[s,\tau]} |X(t;s,\xi) - \xi|^{\bar p} \;\Big|\; \mathcal{F}_s\right] \le C\left(1+|\xi|^{\bar p}\right)\left\{\mathbb{E}\left[|\tau - s|^{\frac{p\bar p}{2(p-\bar p)}} \;\Big|\; \mathcal{F}_s\right]\right\}^{\frac{p-\bar p}{p}}, \\ &\qquad\qquad\qquad \forall (s,\xi) \in \mathcal{D}^p, \quad \tau \in \mathcal{S}[s,T], \quad \bar p \in [1,p), \end{aligned} \tag{2.8}$$

$$\begin{aligned} &\mathbb{E}\left[\sup_{t\in[s\vee\bar s,T]} |X(t;s,\xi) - X(t;\bar s,\xi)|^{\bar p} \;\Big|\; \mathcal{F}_{s\wedge\bar s}\right] \le C(1+|\xi|^{\bar p})\left\{\mathbb{E}\left[|s-\bar s|^{\frac{p\bar p}{2(p-\bar p)}} \;\Big|\; \mathcal{F}_{s\wedge\bar s}\right]\right\}^{\frac{p-\bar p}{p}}, \\ &\qquad\qquad\qquad \forall s, \bar s \in \mathcal{S}[0,T], \quad \xi \in \mathcal{X}^p_{s\wedge\bar s}, \quad \bar p \in [1,p), \end{aligned} \tag{2.9}$$

hereafter, $C > 0$ *represents a generic constant which can be different from line to line.*

A simple consequence of the above is that

$$(t, X(t;s,\xi)) \in \mathcal{D}^p, \qquad t \in \mathcal{S}[s,T], \qquad \forall (s,\xi) \in \mathcal{D}^p. \tag{2.10}$$

We also note that if both $s, \bar s \in [0,T]$ are deterministic, then

$$\mathbb{E}\left[\sup_{t\in[s\vee\bar s,T]} |X(t;s,\xi) - X(t;\bar s,\xi)|^p \;\Big|\; \mathcal{F}_{s\wedge\bar s}\right] \le C(1+|\xi|^p)|s-\bar s|^{\frac{p}{2}} \qquad \forall \xi \in \mathcal{X}^p_{s\wedge\bar s}. \tag{2.11}$$

The following proposition collects some basic results concerning value function $V(\cdot\,,\cdot)$.

Proposition 2.2. *Let* (H1)–(H2) *hold. Then*

(i) *For any* $(s,\xi)\in\mathcal{D}^p$ *and* $\tau\in\mathcal{S}[s,T]$, $J_{s,\xi}(\tau)$ *is a well-defined* $\mathcal{F}_s$*-measurable random variable. Moreover, there exists a* $\bar\tau(s,\xi)\in\mathcal{S}[s,T]$ *such that*

$$V(s,\xi)\equiv\inf_{\tau\in\mathcal{S}[s,T]}J_{s,\xi}(\tau)=J_{s,\xi}(\bar\tau(s,\xi)).\tag{2.12}$$

Consequently, for any $(s,\xi)\in\mathcal{D}^p$, $V(s,\xi)$ *is* $\mathcal{F}_s$*-measurable.*

(ii) *Value function* $V(\cdot\,,\cdot)$ *satisfies the following:*

$$|V(s,\xi)|\le C(1+|\xi|)\qquad\forall(s,\xi)\in\mathcal{D}^p,\tag{2.13}$$

$$\left|V(s,\xi)-V\left(s,\bar\xi\right)\right|\le C\left|\xi-\bar\xi\right|\qquad\forall(s,\xi),\left(s,\bar\xi\right)\in\mathcal{D}^p,\tag{2.14}$$

and when $p>1$,

$$\begin{aligned}\left|\mathbb{E}\left[V(s,\xi)-V(\bar s,\xi)\mid\mathcal{F}_{s\wedge\bar s}\right]\right|\le C(1+|\xi|)\Bigg\{&\left[\mathbb{E}\left(|s-\bar s|^{\frac{p}{2(p-1)}}\mid\mathcal{F}_{s\wedge\bar s}\right)\right]^{\frac{p-1}{p}}\\&+\mathbb{E}\left[\rho(|s-\bar s|)+|s-\bar s|\mid\mathcal{F}_{s\wedge\bar s}\right]\Bigg\},\quad\forall s,\bar s\in\mathcal{S}[0,T],\ \xi\in\mathcal{X}^p_{s\wedge\bar s}.\end{aligned}\tag{2.15}$$

(iii) *For any* $s\in\mathcal{S}[0,T)$ *and any* $\varphi(\cdot)\in L^1_{\mathbb{F}}(\Omega;C([0,T];\mathbb{R}^n))$, *map* $t\mapsto V(t,\varphi(t))$ *is* $\mathbb{F}$*-adapted on* $[s,T]$. *In particular, for any* $(s,\xi)\in\mathcal{D}^p$, *map* $t\mapsto V(t,X(t;s,\xi))$ *is* $\mathbb{F}$*-adapted.*

Proof. (i) By Proposition 2.1 and (H2), we see that, for any fixed $(s,\xi)\in\mathcal{D}^p$,

$$|J_{s,\xi}(\tau)|\le\mathbb{E}\left[\int_s^\tau|g(r,X(r;s,\xi))|dr+|h(\tau,X(\tau;s,\xi))|\mid\mathcal{F}_s\right]\le C(1+|\xi|),\quad\tau\in\mathcal{S}[s,T].$$

Hence, $J_{s,\xi}(\tau)$ is a well-defined $\mathcal{F}_s$-measurable random variable. Next, it is clear that $t\mapsto J_{s,\xi}(t)$ is continuous. Thus, by Theorem D.12 of [14] (see also [8]), with a minor modification, we have the existence of an optimal stopping time $\bar\tau(s,\xi)$ for Problem (S).

(ii) For any $(s,\xi),(s,\bar\xi)\in\mathcal{D}^p$, by (H2) and Proposition 2.1, we can get

$$\begin{aligned}|J_{s,\xi}(\theta)-J_{s,\bar\xi}(\theta)|&\le\mathbb{E}\Bigg[\int_s^\theta\left|g(r,X(r;s,\xi))-g\left(r,X\left(r;s,\bar\xi\right)\right)\right|dr\\&\qquad\qquad+\left|h(\theta,X(\theta;s,\xi))-h\left(\theta,X\left(\theta;s,\bar\xi\right)\right)\right|\mid\mathcal{F}_s\Bigg]\\&\le C\mathbb{E}\left[\sup_{t\in[s,\theta]}\left|X(t;s,\xi)-X\left(t;s,\bar\xi\right)\right|\mid\mathcal{F}_s\right]\le C\left|\xi-\bar\xi\right|\quad\forall\theta\in\mathcal{S}[s,T],\end{aligned}\tag{2.16}$$

with $C>0$ being an absolute constant. Hence, (2.14) follows. Next, let $s,\bar s\in\mathcal{S}[0,T]$,

$\xi \in \mathcal{X}^p_{s\wedge\bar s}$, and $\theta \in \mathcal{S}[s\wedge\bar s, T]$. Observe the following:

$$\begin{aligned}
&\left|\mathbb{E}\left[J_{s,\xi}(s\vee\theta) - J_{\bar s,\xi}(\bar s\vee\theta) \,\middle|\, \mathcal{F}_{s\wedge\bar s}\right]\right|\\
&= \Bigg|\mathbb{E}\Bigg[h(s\vee\theta, X(s\vee\theta; s,\xi)) - h(\bar s\vee\theta, X(\bar s\vee\theta;\bar s,\xi))\\
&\qquad + \int_s^{s\vee\theta} g(t,X(t;s,\xi))dt - \int_{\bar s}^{\bar s\vee\theta} g(t,X(t;\bar s,\xi))dt \,\Big|\, \mathcal{F}_{s\wedge\bar s}\Bigg]\Bigg|\\
&= \Bigg|\mathbb{E}\Bigg\{ I_{(s\vee\bar s\le\theta)}\left[h(\theta,X(\theta;s,\xi)) - h(\theta,X(\theta;\bar s,\xi))\right]\\
&\qquad + I_{(s<\theta<\bar s)}\left[h(\theta,X(\theta;s,\xi)) - h(\bar s,\xi)\right] + I_{(\bar s<\theta<s)}\left[h(s,\xi) - h(\theta,X(\theta;\bar s,\xi))\right]\\
&\qquad + I_{(s\vee\bar s\le\theta)}\Bigg[\int_s^{s\vee\bar s} g(t,X(t;s,\xi))dt - \int_{\bar s}^{s\vee\bar s} g(t,X(t;\bar s,\xi))dt\\
&\qquad + \int_{s\vee\bar s}^{\theta}\left(g(t,X(t;s,\xi))dt - g(t,X(t;\bar s,\xi))\right)dt\Bigg]\\
&\qquad + I_{(s<\theta<\bar s)}\int_s^\theta g(t,X(t;s,\xi))dt - I_{(\bar s<\theta<s)}\int_{\bar s}^\theta g(t,X(t;\bar s,\xi))dt \,\Big|\, \mathcal{F}_{s\wedge\bar s}\Bigg\}\Bigg|\\
&\le \mathbb{E}\Bigg\{ I_{(s\vee\bar s\le\theta)}L|X(\theta;s,\xi) - X(\theta;\bar s,\xi)| + I_{(s<\theta<\bar s)}\left[L|X(\theta;s,\xi)-\xi| + (1+|\xi|)\rho(|s-\bar s|)\right]\\
&\quad + I_{(\bar s<\theta<s)}\left[L|X(\theta;\bar s,\xi)-\xi| + (1+|\xi|)\rho(|s-\bar s|)\right]\\
&\qquad + C\left(1 + \sup_{t\in[s,T]}|X(t;s,\xi)| + \sup_{t\in[\bar s,T]}|X(t;\bar s,\xi)|\right)|s-\bar s|\\
&\qquad + I_{(s\vee\bar s\le\theta)}\int_{s\vee\bar s}^\theta L|X(t;s,\xi) - X(t;\bar s,\xi)|dt \,\Big|\, \mathcal{F}_{s\wedge\bar s}\Bigg\}\\
&\le C\Bigg\{(1+|\xi|)\left[\mathbb{E}\left(|s-\bar s|^{\frac{p}{2(p-1)}} \,\Big|\, \mathcal{F}_{s\wedge\bar s}\right)\right]^{\frac{p-1}{p}} + (1+|\xi|)\mathbb{E}\left[\rho(|s-\bar s|) \,\middle|\, \mathcal{F}_{s\wedge\bar s}\right]\\
&\quad + (1+|\xi|)\mathbb{E}\left[|s-\bar s| \,\middle|\, \mathcal{F}_{s\wedge\bar s}\right] + (1+|\xi|)\left[\mathbb{E}\left(|s-\bar s|^{\frac{p}{2(p-1)}} \,\Big|\, \mathcal{F}_{s\wedge\bar s}\right)\right]^{\frac{p-1}{p}}\Bigg\}\\
&\le C(1+|\xi|)\left\{\left[\mathbb{E}\left(|s-\bar s|^{\frac{p}{2(p-1)}} \,\Big|\, \mathcal{F}_{s\wedge\bar s}\right)\right]^{\frac{p-1}{p}} + \mathbb{E}\left[\rho(|s-\bar s|) + |s-\bar s| \,\middle|\, \mathcal{F}_{s\wedge\bar s}\right]\right\}.
\end{aligned}\tag{2.17}$$

Hence, taking $\theta = \bar\tau(\bar s,\xi)$, we obtain (note $\bar\tau(\bar s,\xi) \ge \bar s$)

$$\begin{aligned}
&\mathbb{E}\left[V(s,\xi) - V(\bar s,\xi) \,\middle|\, \mathcal{F}_{s\wedge\bar s}\right] \le \mathbb{E}\left[J_{s,\xi}(s\vee\bar\tau(\bar s,\xi)) - J_{\bar s,\xi}(\bar\tau(\bar s,\xi)) \,\middle|\, \mathcal{F}_{s\wedge\bar s}\right]\\
&\le C(1+|\xi|)\left\{\left[\mathbb{E}\left(|s-\bar s|^{\frac{p}{2(p-1)}} \,\Big|\, \mathcal{F}_{s\wedge\bar s}\right)\right]^{\frac{p-1}{p}} + \mathbb{E}\left[\rho(|s-\bar s|) + |s-\bar s| \,\middle|\, \mathcal{F}_{s\wedge\bar s}\right]\right\}.
\end{aligned}\tag{2.18}$$

Exchanging the roles of s and $\bar s$, we obtain (2.15).

(iii) is clear. □

3. Principle of optimality and BSPDVI. We now would like to formally derive the equation that value function $V(\cdot\,,\cdot)$ should satisfy. To this end, we first state the following principle of optimality.

THEOREM 3.1. *Let* (H1)–(H2) *hold.*

(i) *For any* $(s,\xi)\in\mathcal{D}^p$,

$$V(s,\xi)\le h(s,\xi) \qquad a.s., \tag{3.1}$$

and

$$V(s,\xi)\le \inf_{\tau\in\mathcal{T}[s,T]}\mathbb{E}\left[\int_s^\tau g(r,X(r;s,\xi))dr+V(\tau,X(\tau;s,\xi))\big|\mathcal{F}_s\right] \quad a.s. \tag{3.2}$$

(ii) *For any* $(s,\xi)\in\mathcal{D}^p$, *if* $\bar\theta\in\mathcal{S}[s,T]$ *is an optimal stopping time of Problem (S) for the initial point* (s,ξ), *then*

$$V\left(\bar\theta,X\left(\bar\theta;s,\xi\right)\right)=h\left(\bar\theta,X\left(\bar\theta;s,\xi\right)\right) \qquad a.s. \tag{3.3}$$

Hence, the following is the smallest optimal stopping time of Problem (S) corresponding to (s,ξ):

$$\bar\tau(s,\xi)=\inf\left\{t\in[s,T]\;\big|\;V(t,X(t;s,\xi))=h(t,X(t;s,\xi))\right\}. \tag{3.4}$$

Moreover,

$$\mathbb{P}\left(\{\bar\tau(s,\xi)>s\}\Delta\{V(s,\xi)<h(s,\xi)\}\right)=0, \tag{3.5}$$

where $A\Delta B=(A\setminus B)\cup(B\setminus A)$, *for any* $A,B\in\mathcal{F}$, *and*

$$\begin{aligned}V(\theta,X(\theta;s,\xi))&=\mathbb{E}\left[\int_\theta^\tau g(r,X(r;s,\xi))dr+V(\tau,X(\tau;s,\xi))\big|\mathcal{F}_\theta\right],\\ &\qquad\forall\theta\in\mathcal{S}[s,\bar\tau(s,\xi)],\quad\tau\in\mathcal{S}[\theta,\bar\tau(s,\xi)],\quad a.s.\end{aligned} \tag{3.6}$$

The above results are basically known (see [8]). For the readers's convenience, we sketch a proof in the appendix.

Note that (3.5) tells us the following: Up to a $\mathbb{P}$-null set, one has

$$\{\bar\tau(s,\xi)>s\}=\{V(s,\xi)<h(s,\xi)\}. \tag{3.7}$$

Consequently, up to a $\mathbb{P}$-null set, the following holds:

$$\{\bar\tau(s,\xi)=s\}=\{V(s,\xi)=h(s,\xi)\}. \tag{3.8}$$

On the other hand, (3.2) implies that

$$\begin{aligned}&V(\theta,X(\theta;s,\xi))+\int_s^\theta g(r,X(r;s,\xi))dr\\ &\qquad\le\mathbb{E}\left[V(\tau,X(\tau;s,\xi))+\int_s^\tau g(r,X(r;s,\xi))dr\big|\mathcal{F}_\theta\right],\\ &\qquad\qquad\forall\theta\in\mathcal{S}[s,T],\ \tau\in\mathcal{S}[\theta,T].\end{aligned} \tag{3.9}$$

This means that

$$\theta\mapsto V(\theta,X(\theta;s,\xi))+\int_s^\theta g(r,X(r;s,\xi))dr$$

is an $\mathbb{F}$-submartingale on $[s,T]$. Likewise, (3.6) implies that

$$
\begin{aligned}
&V(\theta, X(\theta; s,\xi)) + \int_s^\theta g(r, X(r;s,\xi))dr \\
&\qquad = \mathbb{E}\left[V(\tau, X(\tau; s,\xi)) + \int_s^\tau g(r, X(r;s,\xi))dr \Big| \mathcal{F}_\theta \right], \\
&\qquad\qquad \forall \theta \in \mathcal{S}[s, \bar\tau(s,\xi)],\ \tau \in \mathcal{S}[\theta, \bar\tau(s,\xi)],
\end{aligned} \tag{3.10}
$$

which means that

$$
\theta \mapsto V(\theta, X(\theta; s,\xi)) + \int_s^\theta g(r, X(r;s,\xi))dr
$$

is an $\mathbb{F}$-martingale on $[s, \bar\tau(s,\xi)]$.

Next, we would like to derive the HJB equation for value function $V(\cdot\,,\cdot)$. To this end, let us first make a convention: for any differentiable map $f : \mathbb{R}^n \to \mathbb{R}^m$, with $m > 1$, gradient $f_x : \mathbb{R}^n \to \mathbb{R}^{m\times n}$, and for $m = 1$, $f_x : \mathbb{R}^n \to \mathbb{R}^n$. Now, we recall a special case of Itô–Kunita's formula (see [15], [21]).

Theorem 3.2. *Let $F : [0,T] \times \mathbb{R}^n \times \Omega \to \mathbb{R}$ satisfy the following:*

(1) *$(t,x) \mapsto F(t,x,\omega)$ is continuous a.s.;*

(2) *$x \mapsto F(t,x,\omega)$ is C^2 for each $t \in [0,T]$ a.s.;*

(3) *For each $x \in \mathbb{R}^n$, $t \mapsto F(t,x,\cdot)$ is a continuous semimartingale, with*

$$
F(t,x) = F(0,x) + \int_0^t q^0(r,x)dr + \int_0^t \langle\, q(r,x), dW(r)\,\rangle, \qquad (t,x) \in [0,T] \times \mathbb{R}^n
$$

for some $q^0(\cdot)$ and $q(\cdot)$ satisfying the following: For each $x \in \mathbb{R}^n$, $t \mapsto (q^0(t,x), q(t,x))$ is $\mathbb{F}$-adapted, taking values in $\mathbb{R}\times\mathbb{R}^d$, and for almost all $(t,\omega) \in [0,T]\times\Omega$, $x \mapsto q(t,x)$ is C^1. Then

$$
\begin{aligned}
F(t,X(t)) = F(0,X(0)) &+ \textstyle\int_0^t \Big\{ q^0(r,X(r)) + \langle\, b(r,X(r)), F_x(r,X(r))\,\rangle \\
&+ \frac{1}{2}\mathrm{tr}\left[\sigma(r,X(r))\sigma(r,X(r))^T F_{xx}(r,X(r))\right] + \mathrm{tr}\left[\sigma(r,X(r)) q_x(r,X(r))\right] \Big\} dr \\
&+ \int_0^t \left\langle q(r,X(r)) + \sigma(r,X(r))^T F_x(r,X(r)), dW(r) \right\rangle .
\end{aligned} \tag{3.11}
$$

According to our convention, q_x is taking values in $\mathbb{R}^{d\times n}$, and F_x is taking values in $\mathbb{R}^n$. Now, for any $(s,\xi) \in \mathcal{D}^p$, suppose $\bar\tau(s,\xi)$ is the corresponding minimum optimal stopping time. Suppose value function $V(\cdot\,,\cdot)$ admits the following representation:

$$
V(t,x) = V(s,x) + \int_s^t q^0(r,x)dr + \int_s^t \langle\, q(r,x), dW(r)\,\rangle, \qquad (t,x) \in [s,T] \times \mathbb{R}^n,
$$

with $q^0(\cdot)$ and $q(\cdot)$ being undetermined. Then, by Itô–Kunita's formula, for any $t \in \mathcal{S}[s,T]$,

$$
\begin{aligned}
V(t, X(t;s,\xi)) = V(s,\xi) &+ \int_s^t \Big\{ q^0(r, X(r;s,\xi)) + \langle\, b(r,X(r;s,\xi)), V_x(r,X(r;s,\xi))\,\rangle \\
&+ \frac{1}{2}\mathrm{tr}\left[\sigma(r,X(r;s,\xi))\sigma(r,X(r;s,\xi))^T V_{xx}(r,X(r;s,\xi))\right] \\
&+ \mathrm{tr}\left[\sigma(r,X(r;s,\xi)) q_x(r,X(r;s,\xi))\right] \Big\} dr \\
&+ \int_s^t \left\langle q(r,X(r;s,\xi)) + \sigma(r,X(r;s,\xi))^T V_x(r,X(r;s,\xi)), dW(r) \right\rangle .
\end{aligned} \tag{3.12}
$$

Hence, by (3.2), we have

$$\begin{aligned}
0 &\le \mathbb{E}\left[\int_s^t g(r,X(r;s,\xi))dr + V(t,X(t;s,\xi)) - V(s,\xi)|\mathcal{F}_s\right]\\
&= \mathbb{E}\Bigg[\int_s^t \Big\{g(r,X(r;s,\xi)) + q^0(r,X(r;s,\xi)) + \langle\, b(r,X(r;s,\xi)), V_x(r,X(r;s,\xi))\,\rangle\\
&\quad + \frac{1}{2}\text{tr }\left[\sigma(r,X(r;s,\xi))\sigma(r,X(r;s,\xi))^T V_{xx}(r,X(r;s,\xi))\right]\\
&\quad + \text{tr }[\sigma(r,X(r;s,\xi))q_x(r,X(r;s,\xi))]\Big\}dr \;\Big|\; \mathcal{F}_s\Bigg].
\end{aligned} \tag{3.13}$$

Dividing it by $(t-s)$ and sending $t \to s$, we obtain

$$\begin{aligned}
0 &\le g(s,\xi) + q^0(s,\xi) + V_x(s,\xi)b(s,\xi) + \frac{1}{2}\text{tr }\left[\sigma(s,\xi)\sigma(s,\xi)^T V_{xx}(s,\xi)\right]\\
&\quad + \text{tr }[\sigma(s,\xi)q_x(s,\xi)] \qquad \text{a.s.,} \quad \forall(s,\xi)\in\mathcal{D}^p.
\end{aligned} \tag{3.14}$$

On the other hand, on the set $\{V(s,\xi) < h(s,\xi)\}$, one has $\bar\tau(s,\xi) > s$, and

$$\theta \mapsto V(\theta, X(\theta;s,\xi)) + \int_s^\theta g(t,X(t;s,\xi))dt$$

is a martingale on $[s,\bar\tau(s,\xi))$. Hence, it is necessary that

$$\begin{aligned}
0 &= g(s,\xi) + q^0(s,\xi) + \left\langle b(s,\xi), V_x(s,\xi)^T\right\rangle + \frac{1}{2}\text{tr }\left[\sigma(s,\xi)\sigma(s,\xi)^T V_{xx}(s,\xi)\right]\\
&\quad + \text{tr }[\sigma(s,\xi)q_x(s,\xi)] \qquad \text{a.s. on } \{V(s,\xi) < h(s,\xi)\}, \quad \forall(s,\xi)\in\mathcal{D}^p.
\end{aligned} \tag{3.15}$$

Therefore, it is reasonable to require that

$$\begin{cases}
q^0(r,x) \ge -\frac{1}{2}\text{tr }\left[\sigma(r,x)\sigma(r,x)^T V_{xx}(r,x)\right] - \langle\, b(r,x), V_x(r,x)\,\rangle\\
\qquad\qquad - \text{tr }[\sigma(r,x)q_x(r,x)] - g(r,x) \qquad \text{a.s.,} \quad (r,x)\in[0,T]\times\mathbb{R}^n,\\
q^0(r,x) = -\frac{1}{2}\text{tr }\left[\sigma(r,x)\sigma(r,x)^T V_{xx}(r,x)\right] - \langle\, b(r,x), V_x(r,x)\,\rangle\\
\qquad\qquad - \text{tr }[\sigma(r,x)q_x(r,x)] - g(r,x) \quad \text{as on } \{V(r,x) < h(r,x)\},\\
\qquad\qquad\qquad\qquad\qquad\qquad\qquad (r,x)\in[0,T]\times\mathbb{R}^n.
\end{cases} \tag{3.16}$$

If we let $\beta : \mathbb{R} \to [0,+\infty]$ be a monotone graph defined by

$$\beta(\rho) = \begin{cases}
\phi, & \rho > 0,\\
[0,+\infty), & \rho = 0,\\
0, & \rho < 0,
\end{cases} \tag{3.17}$$

then we should have

$$\begin{aligned}
q^0(r,x) &\in -\frac{1}{2}\text{tr }\left[\sigma(r,x)\sigma(r,x)^T V_{xx}(r,x)\right] - \langle\, b(r,x), V_x(r,x)\,\rangle - \text{tr }[\sigma(r,x)q_x(r,x)]\\
&\quad + \beta\left(V(r,x) - h(r,x)\right) - g(r,x), \qquad \text{a.s.} \quad (r,x)\in[0,T]\times\mathbb{R}^n,
\end{aligned} \tag{3.18}$$

which is understood as follows:

$$\begin{cases} q^0(r,x) = -\frac{1}{2}\text{tr }\left[\sigma(r,x)\sigma(r,x)^T V_{xx}(r,x)\right] - \langle\, b(r,x), V_x(r,x)\,\rangle - \text{tr }\left[\sigma(r,x) q_x(r,x)\right] \\ \qquad\qquad + \zeta(r,x) - g(r,x) \qquad \text{a.s.} \quad (r,x)\in[0,T]\times\mathbb{R}^n, \\ \zeta(r,x) \in \beta\left(V(r,x) - h(r,x)\right) \qquad \text{a.s.} \quad (r,x)\in[0,T]\times\mathbb{R}^n. \end{cases} \tag{3.19}$$

In the above, $\zeta(\cdot,x)$ is required to be $\mathbb{F}$-adapted. Consequently, we should have

$$\begin{cases} V(t,x) = h(T,x) + \displaystyle\int_t^T \Big\{ \frac{1}{2}\text{tr }\left[\sigma(r,x)\sigma(r,x)^T V_{xx}(r,x)\right] + \langle\, b(r,x), V_x(r,x)\,\rangle \\ \qquad + \text{tr }\left[\sigma(r,x) q_x(r,x)\right] \\ \qquad - \zeta(r,x) + g(r,x) \Big\} dr - \displaystyle\int_t^T \langle\, q(r,x), dW(r)\,\rangle, \qquad t\in[0,T], \quad x\in\mathbb{R}^n, \\ \zeta(t,x) \in \beta\left(V(t,x) - h(t,x)\right), \qquad t\in[0,T], \quad x\in\mathbb{R}^n. \end{cases} \tag{3.20}$$

We call (3.20) a BSPDVI. Note that, in (3.20), the unknown is the triple of $\mathbb{F}$-adapted random fields $(V,q,\zeta) : [0,T]\times\mathbb{R}^n\times\Omega \to \mathbb{R}\times\mathbb{R}^d\times\mathbb{R}$. Note that the last inclusion in (3.20) is equivalent to the following:

$$\begin{cases} V(t,x) - h(t,x) \le 0, \quad \zeta(t,x)\ge 0, \qquad (t,x)\in[0,T]\times\mathbb{R}^n \ \text{ a.s.}, \\ \left[V(t,x) - h(t,x)\right]\zeta(t,x) = 0, \qquad (t,x)\in[0,T]\times\mathbb{R}^n \ \text{a.s.} \end{cases} \tag{3.21}$$

4. Adapted solutions. In this section, we will introduce notions of adapted solutions for BSPDVI (3.20) and will carry out some preliminary studies. To begin with, let us make a little preparation.

By a multi-index α, we mean $\alpha = (\alpha_1,\ldots,\alpha_n)$, with each α_i being nonnegative integers, and we define $|\alpha| = \sum_{i=1}^n \alpha_i$. We write $x = (x_1,\ldots,x_n)$ for any generic point in $\mathbb{R}^n$. For any multi-index $\alpha \equiv (\alpha_1,\ldots,\alpha_n)$ and any smooth function $f(\cdot)$, denote

$$\partial^\alpha f(x) = \partial_{x_1}^{\alpha_1}\cdots\partial_{x_n}^{\alpha_n} f(x). \tag{4.1}$$

For any domain $G\subseteq\mathbb{R}^n$ (G is allowed to be $\mathbb{R}^n$), let $C^k(G,\mathbb{R})$ be the set of all functions $f : G\to\mathbb{R}$ such that

$$\sup_{x\in G,\,|\alpha|\le k} |\partial^\alpha f(x)| < \infty. \tag{4.2}$$

We may similarly define spaces $C^k(G;\mathbb{R}^n)$ and $C^k(G;\mathbb{R}^{n\times d})$, etc. Clearly, these are Banach spaces. Next, we let $W^{m,p}(G;\mathbb{R})$ be the usual Sobolev space of all functions $f(\cdot)$ such that

$$\|f(\cdot)\|^p_{W^{m,p}(G)} \equiv \sum_{|\alpha|\le m}\int_G |\partial^\alpha f(x)|^p dx < \infty, \tag{4.3}$$

and $H^m(\mathbb{R}^n) = W^{m,2}(\mathbb{R}^n)$. For any Banach space B, let $L^\infty_{\mathbb{F}}(0,T;B)$ be the space of all bounded $\mathbb{F}$-progressively measurable maps $f : [0,T]\times\Omega\to B$, with the norm

$$\|f(\cdot)\|_{L^\infty_{\mathbb{F}}(0,T;B)} = \operatorname*{esssup}_{(t,\omega)\in[0,T]\times\Omega} \|f(t,\omega)\|_B. \tag{4.4}$$

Here, B could be $C^k(G;\mathbb{R}^n)$, say. Similarly, we let $C_{\mathbb{F}}(0,T;B)$ be the space of all B-valued $\mathbb{F}$-adapted continuous processes, which is a closed subspace of $L^\infty_{\mathbb{F}}(0,T;B)$.

We now introduce the following definition.

Definition 4.1. (i) *A triple of random fields* (V, q, ζ) *is called an adapted strong solution of* (3.20) *if for each* $x \in \mathbb{R}^n$, $t \mapsto (V(t,x), q(t,x), \zeta(t,x))$ *is* $\mathbb{F}$*-adapted and for almost all* $(t,x,\omega) \in [0,T] \times \mathbb{R}^n \times \Omega$, $x \mapsto V(t,x,\omega)$ *is twice differentiable,* $x \mapsto q(t,x,\omega)$ *is once differentiable such that* (3.20) *is satisfied for almost all* $(x,\omega) \in \mathbb{R}^n \times \Omega$.

(ii) *An adapted strong solution* (V, q, ζ) *of* (3.20) *is called an adapted classical solution of* (3.20) *if for almost all* $(t,\omega) \in [0,T] \times \Omega$, $x \mapsto V(t,x,\omega)$ *is* C^2, $x \mapsto q(t,x,\omega)$ *is* C^1, *and* $x \mapsto \zeta(t,x)$ *is continuous.*

Once we have the well-posedness of our BSPDVI (which will be treated in the next section), it is natural to ask if solution $V(\cdot\,,\cdot)$ to the BSPDVI has anything to do with our Problem (S)? The following result answers this question: Under appropriate conditions, the solution of BSPDVI (3.20) coincides with the value function of Problem (S), via which an optimal stopping time can be identified.

Theorem 4.2. *Let* (H1)–(H2) *hold. Suppose* (V, q, ζ) *is an adapted classical solution to BSPDVI* (3.20). *Then* $V(\cdot\,,\cdot)$ *is the value function of Problem (S). Consequently, part* $V(\cdot\,,\cdot)$ *of adapted classical solution* (V, q, ζ) *to* (3.20) *is unique. Moreover, the following gives the smallest optimal stopping time of Problem (S):*

$$\bar\tau(s,\xi) = \inf\left\{ t \in [s,T] \;\middle|\; V(t, X(t;s,\xi)) = h(t, X(t;s,\xi)) \right\}. \tag{4.5}$$

Proof. Let $(s,\xi) \in \mathcal{D}^p$, and define $\bar\tau(s,\xi)$ by (4.5). By the Itô–Kunita formula, together with BSPDVI (3.20), we have

$$V(s,\xi) = \mathbb{E}\left[\int_s^\tau g(r, X(r;s,\xi))dr + V(\tau, X(\tau;s,\xi)) \big| \mathcal{F}_s \right] \quad \forall \tau \in \mathcal{S}[s, \bar\tau(s,\xi)]. \tag{4.6}$$

Hence, taking $\tau = \bar\tau(s,\xi)$, we have

$$\begin{aligned} V(s,\xi) &= \mathbb{E}\left[V(\bar\tau(s,\xi), X(\bar\tau(s,\xi);s,\xi)) + \int_s^{\bar\tau(s,\xi)} g(r, X(r;s,\xi))dr \;\big|\; \mathcal{F}_s \right] \\ &= \mathbb{E}\left[h(\bar\tau(s,\xi), X(\bar\tau(s,\xi);s,\xi)) + \int_s^{\bar\tau(s,\xi)} g(r, X(r;s,\xi))dr \;\big|\; \mathcal{F}_s \right] = J_{s,\xi}(\bar\tau(s,\xi)). \end{aligned} \tag{4.7}$$

This means that $\bar\tau(s,\xi)$ is an optimal stopping time for our Problem (S). From the above, we further conclude that part $V(\cdot\,,\cdot)$ of adapted solution (V, q, ζ) to BSPDVI (3.20) is unique, and from (4.5), $\bar\tau(s,\xi)$ has to be the smallest optimal stopping time (noting (ii) of Theorem 3.1). ▯

Next, we would like to make a reduction which will be very useful below. To this end, let

$$h(t,x) = h(0,x) + \int_0^t \mu^0(r,x)dr + \int_0^t \langle\, \mu(r,x), dW(r)\,\rangle, \qquad t \in [0,T] \tag{4.8}$$

for some suitable $\mu^0(\cdot)$ and $\mu(\cdot)$. Suppose (V, q, ζ) is an adapted classical solution to BSPDVI (3.20), and all the coefficients have required order of derivatives. We fix $p \ge 2$, and let

$$\begin{cases} \bar V(t,x) = \dfrac{V(t,x) - h(t,x)}{1+|x|^p}, \\ \bar q(t,x) = \dfrac{q(t,x) - \mu(t,x)}{1+|x|^p}. \\ \bar\zeta(t,x) = \dfrac{\zeta(t,x)}{1+|x|^p}. \end{cases} \qquad (t,x) \in [0,T] \times \mathbb{R}^n \ \text{ a.s.}, \tag{4.9}$$

Note that, in the case $x \mapsto V(t,x) - h(t,x)$ grows at most linearly, $x \mapsto \bar V(t,x)$ will be L^2-integrable over $\mathbb{R}^n$. According to the above, one has

$$\begin{cases} V(t,x) = (1+|x|^p)\bar V(t,x) + h(t,x), \\ q(t,x) = (1+|x|^p)\bar q(t,x) + \mu(t,x), \\ \zeta(t,x) = (1+|x|^p)\bar\zeta(t,x). \end{cases} \tag{4.10}$$

Consequently (suppressing (t,x)),

$$\begin{cases} V_x = (1+|x|^p)\,\bar V_x + p|x|^{p-2}\bar V x + h_x, \\ V_{xx} = (1+|x|^p)\,\bar V_{xx} + p|x|^{p-2}\left[\left(x\bar V_x\right)^T + x\bar V_x\right] \\ \qquad + \left[p(p-2)|x|^{p-4}xx^T + p|x|^{p-2}I\right]\bar V + h_{xx}, \\ q_x = (1+|x|^p)\,\bar q_x + p|x|^{p-2}\bar q x^T + \mu_x. \end{cases} \tag{4.11}$$

Hence,

$$\begin{aligned} \bar V(t,x) &\equiv (1+|x|^p)^{-1}\left[V(t,x) - h(t,x)\right] \\ &= (1+|x|^p)^{-1}\Bigg(h(T,x) - h(t,x) + \int_t^T \left\{\frac{1}{2}\mathrm{tr}\,\left[\sigma(r,x)\sigma(r,x)^T V_{xx}(r,x)\right]\right. \\ &\quad \left. + \langle\, b(r,x), V_x(r,x)\,\rangle + \mathrm{tr}\,\left[\sigma(r,x)q_x(r,x)\right] - \zeta(r,x) + g(r,x)\right\}dr \\ &\quad - \int_t^T \langle\, q(r,x), dW(r)\,\rangle\Bigg) \\ &= \int_t^T \frac{\mu^0(r,x)}{1+|x|^p}dr + \int_t^T \left\langle \frac{\mu(r,x)}{1+|x|^p}, dW(r)\right\rangle + \int_t^T \left\{\frac{1}{2}\mathrm{tr}\left[\sigma(r,x)\sigma(r,x)^T\right.\right. \\ &\quad \cdot\left.\left(\bar V_{xx} + \frac{p|x|^{p-2}[x\bar V_x^T + \bar V_x x^T]}{1+|x|^p} + \frac{p|x|^{p-4}[(p-2)xx^T + |x|^2 I]\bar V + h_{xx}}{1+|x|^p}\right)\right] \\ &\quad + \left\langle b(r,x), \bar V_x + \frac{p|x|^{p-2}\bar V x + h_x}{1+|x|^p}\right\rangle + \mathrm{tr}\,\left[\sigma(r,x)\left(\bar q_x + \frac{p|x|^{p-2}\bar q x^T + \mu_x}{1+|x|^p}\right)\right] \\ &\quad \left. - \frac{\zeta(r,x) - g(r,x)}{1+|x|^p}\right\}dr - \int_t^T \left\langle \frac{q(r,x)}{1+|x|^p}, dW(s)\right\rangle \\ &= \int_t^T \left\{\frac{1}{2}\mathrm{tr}\,\left[\sigma(r,x)\sigma(r,x)^T\bar V_{xx}\right] + \left\langle b(r,x) + \frac{p|x|^{p-2}\sigma(r,x)\sigma(r,x)^T x}{1+|x|^p}, \bar V_x\right\rangle\right. \\ &\quad + \frac{p|x|^{p-4}[(p-2)|\sigma(r,x)^T x|^2 + |x|^2|\sigma(r,x)|^2] + 2p|x|^{p-2}\langle\, b(r,x), x\,\rangle}{2(1+|x|^p)}\bar V \\ &\quad + \mathrm{tr}\,\left[\sigma(r,x)\bar q_x\right] + \left\langle \frac{p|x|^{p-2}\sigma(r,x)^T x}{1+|x|^p}, \bar q\right\rangle - \bar\zeta \\ &\quad + (1+|x|^p)^{-1}\left[\frac{1}{2}\mathrm{tr}\,\big(\sigma(r,x)\sigma(r,x)^T h_{xx}(r,x)\big) + \langle\, b(r,x), h_x(r,x)\,\rangle\right. \\ &\quad \left.\left. + \mathrm{tr}\,(\sigma(r,x)\mu_x(r,x)) + \mu^0(r,x) + g(r,x)\right]\right\}dr - \int_s^t \langle\, \bar q(r,x), dW(s)\,\rangle \\ &\equiv \int_s^t \left\{\frac{1}{2}\mathrm{tr}\,\left[\sigma(r,x)\sigma(r,x)^T\bar V_{xx}(r,x)\right] + \left\langle \widetilde b(r,x), \bar V_x(r,x)\right\rangle + \widetilde b^0(r,x)\bar V(r,x)\right. \\ &\quad \left. + \mathrm{tr}\,[\sigma(r,x)\bar q_x(r,x)] + \langle\, \widetilde\sigma^0(r,x), \bar q(r,x)\,\rangle - \bar\zeta(r,x) + \widetilde g(r,x)\right\}dr \\ &\quad - \int_s^t \langle\, \bar q(r,x), dW(s)\,\rangle, \end{aligned} \tag{4.12}$$

with

$$\begin{cases} \widetilde{b}(r,x) = b(r,x) + \dfrac{p|x|^{p-2}\sigma(r,x)\sigma(r,x)^T x}{1+|x|^p}, \\ \widetilde{b}^0(r,x) = \dfrac{p|x|^{p-4}[(p-2)|\sigma(r,x)^T x|^2 + |x|^2|\sigma(r,x)|^2] + 2p|x|^{p-2}\left\langle b(r,x), x \right\rangle}{2(1+|x|^p)}, \\ \widetilde{\sigma}^0(r,x) = \dfrac{p|x|^{p-2}\sigma(r,x)^T x}{1+|x|^p}, \\ \widetilde{g}(r,x) = (1+|x|^p)^{-1}\Big[\dfrac{1}{2}\mathrm{tr}\,\left(\sigma(r,x)\sigma(r,x)^T h_{xx}(r,x)\right) + \left\langle b(r,x), h_x(r,x) \right\rangle \\ \qquad + \mathrm{tr}\,\left(\sigma(r,x)\mu_x(r,x)\right) + \mu^0(r,x) + g(r,x)\Big]. \end{cases} \tag{4.13}$$

Note that, by the definition of β, we see that

$$\zeta(t,x) \in \beta\left(V(t,x) - h(t,x)\right), \qquad (t,x) \in [0,T]\times\mathbb{R}^n \quad \text{a.s.} \tag{4.14}$$

is equivalent to

$$\bar{\zeta}(t,x) \in \beta\left(\bar{V}(t,x)\right), \qquad (t,x) \in [0,T]\times\mathbb{R}^n \quad \text{a.s.} \tag{4.15}$$

Next, we have

$$\begin{cases} \mathrm{tr}\,\left(\sigma\sigma^T \bar{V}_{xx}\right) = \nabla\cdot\left(\sigma\sigma^T\bar{V}_x\right) - \left\langle \nabla\cdot\left(\sigma\sigma^T\right), \bar{V}_x\right\rangle, \\ \mathrm{tr}\,\left(\sigma\bar{q}_x\right) = \nabla\cdot\left(\sigma\bar{q}\right) - \left\langle \nabla\cdot\sigma, \bar{q}\right\rangle, \end{cases} \tag{4.16}$$

where (with $\sigma = (\sigma_1,\dots,\sigma_d)$, each σ_i takes values in $\mathbb{R}^n$)

$$\nabla\cdot\sigma = \left(\nabla\cdot\sigma_1,\dots,\nabla\cdot\sigma_d\right)^T \tag{4.17}$$

and

$$\begin{aligned} \nabla\cdot\left(\sigma\sigma^T\right) &= \sum_{k=1}^d \nabla\cdot\left(\sigma_k\sigma_k^T\right) = \sum_{k=1}^d \left(\nabla\cdot\left(\sigma_{1k}\sigma_k\right),\dots,\nabla\cdot\left(\sigma_{nk}\sigma_k\right)\right)^T \\ &= \sum_{k=1}^d \left(\sigma_{1k}(\nabla\cdot\sigma_k),\dots,\sigma_{nk}(\nabla\cdot\sigma_k)\right)^T + \sum_{k=1}^d \left(\sigma_k^T(\sigma_{1k})_x,\dots,\sigma_k^T(\sigma_{nk})_x\right)^T \\ &= \sum_{k=1}^d (\nabla\cdot\sigma_k)\sigma_k + \sum_{k=1}^d (\sigma_k)_x\sigma_k = \sum_{k=1}^d \left[(\nabla\cdot\sigma_k)I + (\sigma_k)_x\right]\sigma_k. \end{aligned} \tag{4.18}$$

Then, we can get

$$\begin{aligned} &\frac{1}{2}\mathrm{tr}\,\left(\sigma\sigma^T\bar{V}_{xx}\right) + \left\langle \widetilde{b}, \bar{V}_x\right\rangle + \widetilde{b}^0\bar{V} + \mathrm{tr}\,\left(\sigma\bar{q}_x\right) + \left\langle\widetilde{\sigma}^0, \bar{q}\right\rangle - \bar{\zeta} + \widetilde{g} \\ &= \frac{1}{2}\nabla\cdot\left(\sigma\sigma^T\bar{V}_x\right) + \left\langle \widetilde{b} - \nabla\cdot\left(\sigma\sigma^T\right), \bar{V}_x\right\rangle + \widetilde{b}^0\bar{V} + \nabla\cdot\left(\sigma\bar{q}\right) + \left\langle\widetilde{\sigma}^0 - \nabla\cdot\sigma, \bar{q}\right\rangle - \bar{\zeta} + \bar{g}. \end{aligned} \tag{4.19}$$

According to the above reduction, we have the following divergence form of our

BSPDVI:

$$
\begin{cases}
\bar V(t,x)=\int_t^T\Big\{\frac12\nabla\cdot\big[\sigma(r,x)\sigma(r,x)^T\bar V_x(r,x)\big]+\big\langle\bar b(r,x),\bar V_x(r,x)\big\rangle+\bar b^0(r,x)\bar V(r,x)\\
\qquad\qquad+\nabla\cdot[\sigma(r,x)\bar q(r,x)]+\langle\,\bar\sigma^0(r,x),\bar q(r,x)\,\rangle-\bar\zeta(r,x)+\bar g(r,x)\Big\}dr\\
\qquad\qquad-\displaystyle\int_t^T\langle\bar q(r,x),dW(r)\rangle\,,\qquad t\in[0,T],\quad x\in\mathbb{R}^n,\\
\bar\zeta(t,x)\in\beta\big(\bar V(t,x)\big)\,,\qquad t\in[0,T],\quad x\in\mathbb{R}^n,
\end{cases}
\tag{4.20}
$$

with

$$
\begin{cases}
\bar b(r,x)=\widetilde b(r,x)-\nabla\cdot\big[\sigma(r,x)\sigma(r,x)^T\big]\\
\qquad\quad=b(r,x)-\nabla\cdot\big[\sigma(r,x)\sigma(r,x)^T\big]+\dfrac{p|x|^{p-2}\sigma(r,x)\sigma(r,x)^Tx}{1+|x|^p}\,,\\
\bar b^0(r,x)=\widetilde b^0(r,x)\\
\qquad\quad=\dfrac{p|x|^{p-4}[(p-2)|\sigma(r,x)^Tx|^2+|x|^2|\sigma(r,x)|^2]+2p|x|^{p-2}\,\langle\,b(r,x),x\,\rangle}{2(1+|x|^p)}\,,\\
\bar\sigma^0(r,x)=\widetilde\sigma^0(r,x)-\nabla\cdot\sigma(r,x)=\dfrac{p|x|^{p-2}\sigma(r,x)^Tx}{1+|x|^p}-\nabla\cdot\sigma(r,x)\,,\\
\bar g(r,x)=\widetilde g(r,x)=(1+|x|^p)^{-1}\bigg[\dfrac12\mathrm{tr}\,\big(\sigma(r,x)\sigma(r,x)^Th_{xx}(r,x)\big)+\langle\,b(r,x),h_x(r,x)\,\rangle\\
\qquad\quad+\,\mathrm{tr}\,(\sigma(r,x)\mu_x(r,x))+\mu^0(r,x)+g(r,x)\bigg].
\end{cases}
\tag{4.21}
$$

In order for the above reduction to be possible and for the purpose of some other further discussions, we introduce the following assumption.

Assumption (H3). For some $k>2+\frac{n}{2}$, maps $b(\cdot)$, $\sigma(\cdot)$, $g(\cdot)$, $\mu_0(\cdot)$, $\mu(\cdot)$, and $h(0,\cdot)$ satisfy the following:

$$
\begin{cases}
b(\cdot)\in L^\infty_{\mathbb{F}}\big(0,T;C^k(\mathbb{R}^n;\mathbb{R}^n)\big)\,,\quad\sigma(\cdot),\mu(\cdot)\in L^\infty_{\mathbb{F}}\big(0,T;C^k(\mathbb{R}^n;\mathbb{R}^{n\times d})\big)\,,\\
g(\cdot),\mu_0(\cdot)\in L^\infty_{\mathbb{F}}\big(0,T;C^k(\mathbb{R}^n;\mathbb{R})\big)\,,\quad h(0,\cdot)\in C^k(\mathbb{R}^n;\mathbb{R}).
\end{cases}
\tag{4.22}
$$

Under (H3), we see that

$$
\begin{cases}
\big|\bar b(t,x)\big|+\big|\bar b^0(t,x)\big|+\big|\bar\sigma^0(t,x)\big|\le C,\\
|\bar g(t,x)|\le C(1+|x|^p)^{-1},
\end{cases}
\qquad (t,x)\in(t,x)\in[0,T]\times\mathbb{R}^n\quad\text{a.s.}
\tag{4.23}
$$

In what follows, we will choose $p>2$, which will lead to

$$
\int_{\mathbb{R}^n}|\bar g(t,x)|dx\le C\qquad\forall t\in[0,T]\text{ a.s.}
\tag{4.24}
$$

We may introduce adapted classical and strong solutions to the divergence form of BSPDVI (4.20), similar to Definition 4.1. On the other hand, let us now introduce the following notion.

Definition 4.3. *A triple*

$$
\big(\bar V,\bar q,\bar\zeta\big)\in L^2_{\mathbb{F}}\big(0,T;H^1(\mathbb{R}^n)\big)\times L^2_{\mathbb{F}}\big(0,T;L^2\big(\mathbb{R}^n;\mathbb{R}^d\big)\big)\times L^2_{\mathbb{F}}\big(0,T;L^2(\mathbb{R}^n)\big)
\tag{4.25}
$$

is called an adapted weak solution of (4.20) *if, for any* $\varphi \in H^1(\mathbb{R}^n)$,

$$\begin{aligned}\int_{\mathbb{R}^n} \bar V(t,x)\varphi(x)dx = \int_t^T\int_{\mathbb{R}^n}\Big\{ &-\Big\langle \frac{1}{2}\sigma(r,x)^T\bar V_x(r,x) + \bar q(r,x), \sigma(r,x)^T\varphi_x(x)\Big\rangle \\ &+[\big\langle \bar b(r,x), \bar V_x(r,x)\big\rangle + \bar b^0(r,x)\bar V(r,x) \\ &+ \big\langle \bar\sigma^0(r,x), \bar q(r,x)\big\rangle + \bar g(r,x) - \bar\zeta(r,x)]\varphi(x)\Big\}dxdr \\ &- \int_t^T \Big\langle \int_{\mathbb{R}^n} \bar q(r,x)\varphi(x)dx, dW(r)\Big\rangle, \qquad t\in[0,T],\end{aligned} \tag{4.26}$$

and

$$\begin{cases} \bar V(t,x) \le 0, \quad \bar\zeta(t,x)\ge 0, \qquad (t,x)\in[0,T]\times\mathbb{R}^n \quad a.s., \\ \bar V(t,x)\bar\zeta(t,x) = 0, \qquad\qquad \text{a.e. } (t,x)\in[0,T]\times\mathbb{R}^n \quad \text{a.s.}\end{cases} \tag{4.27}$$

We point out that any adapted strong solution $(\bar V, \bar q, \bar\zeta)$ of (4.20) must be an adapted weak solution of (4.20). Similar to [16], we can show, by an argument using integration by parts, that an adapted weak solution is an adapted strong (classical) solution if it has the regularity that the latter requires.

5. Well-posedness of the BSPDVI. In this section, we are going to discuss the issue of the well-posedness of BSPVDI. First, we have the following.

THEOREM 5.1. *Suppose* (H3) *holds. Let* $(\bar V, \bar q, \bar\zeta)$ *and* $(\widetilde V, \widetilde q, \widetilde\zeta)$ *be adapted weak solutions to* (4.20)*, with*

$$\begin{cases} \bar V, \widetilde V \in L^2_{\mathbb{F}}\left(0,T; H^1(\mathbb{R}^n)\right), \\ \bar q, \widetilde q \in L^2_{\mathbb{F}}\left(0,T; L^2(\mathbb{R}^n;\mathbb{R}^d)\right), \\ \bar\zeta, \widetilde\zeta \in L^2_{\mathbb{F}}\left(0,T;L^2(\mathbb{R}^n)\right). \end{cases} \tag{5.1}$$

Then

$$\bar V(t,x) = \widetilde V(t,x), \quad \bar q(t,x) = \widetilde q(t,x), \quad \bar\zeta(t,x) = \widetilde\zeta(t,x), \qquad (t,x)\in[0,T]\times\mathbb{R}^n, \text{ a.s.} \tag{5.2}$$

Proof. Suppose $(\bar V, \bar q, \bar\zeta)$ and $(\widetilde V, \widetilde q, \widetilde\zeta)$ are two adapted weak solutions to BSPDVI (4.20). Set

$$\widehat V = \bar V - \widetilde V, \qquad \widehat q = \bar q - \widetilde q, \qquad \widehat\zeta = \bar\zeta - \widetilde\zeta. \tag{5.3}$$

Then $(\widehat V, \widehat q, \widehat\zeta)$ is an adapted weak solution to the following linear BSPDE:

$$\begin{cases} \widehat V(t,x) = \displaystyle\int_t^T \Big\{\frac{1}{2}\nabla\cdot\Big[\sigma(r,x)\sigma(r,x)^T\widehat V_x(r,x)\Big] + \Big\langle \bar b(r,x), \widehat V_x(r,x)\Big\rangle + \bar b^0(r,x)\widehat V(r,x) \\ \qquad\qquad + \nabla\cdot[\sigma(r,x)\widehat q(r,x)] + \big\langle\bar\sigma^0(r,x), \widehat q(r,x)\big\rangle - \widehat\zeta(r,x)\Big\}dr \\ \qquad\qquad - \displaystyle\int_t^T \langle\, \widehat q(r,x), dW(r)\,\rangle, \qquad t\in[0,T], \quad x\in\mathbb{R}^n. \end{cases} \tag{5.4}$$

By Itô's type formula (see [16] for details), we have

$$\begin{aligned}\mathbb{E}\int_{\mathbb{R}^n}\left|\widehat{V}(t,x)\right|^2dx=\mathbb{E}\int_t^T\int_{\mathbb{R}^n}\Big\{-\left|\sigma^T\widehat{V}_x\right|^2+[2\bar{b}^0-(\nabla\cdot\bar{b})]\widehat{V}^2-2\left\langle\widehat{q},\sigma^T\widehat{V}_x\right\rangle\\ +2\left\langle\bar{\sigma}^0,\widehat{q}\right\rangle\widehat{V}-2\widehat{\zeta}\widehat{V}-|\widehat{q}|^2\Big\}dxdr\\ =\mathbb{E}\int_t^T\int_{\mathbb{R}^n}\Big\{-\left|\sigma^T\widehat{V}_x+\widehat{q}-\bar{\sigma}^0\widehat{V}\right|^2+\left|\bar{\sigma}^0\right|^2\widehat{V}^2\\ -2\left\langle\sigma^T\widehat{V}_x,\bar{\sigma}^0\widehat{V}\right\rangle+\ [2\bar{b}^0-\big(\nabla\cdot\bar{b}\big)]\widehat{V}^2-2\widehat{\zeta}\widehat{V}\Big\}dxdr\\ =\mathbb{E}\int_t^T\int_{\mathbb{R}^n}\Big\{-\left|\sigma^T\widehat{V}_x+\widehat{q}-\bar{\sigma}^0\widehat{V}\right|^2\\ +\Big[\left|\bar{\sigma}^0\right|^2+\nabla\cdot\big(\sigma\bar{\sigma}^0\big)+2\bar{b}^0-\big(\nabla\cdot\bar{b}\big)\Big]\widehat{V}^2-2\widehat{\zeta}\widehat{V}\Big\}dxdr.\end{aligned}\tag{5.5}$$

Note that

$$\begin{aligned}\widehat{V}(t,x)\widehat{\zeta}(t,x)&\equiv\left[\bar{V}(t,x)-\widetilde{V}(t,x)\right]\left[\bar{\zeta}(t,x)-\widetilde{\zeta}(t,x)\right]\\ &=-\bar{V}(t,x)\widetilde{\zeta}(t,x)-\widetilde{V}(t,x)\bar{\zeta}(t,x)\geq 0.\end{aligned}\tag{5.6}$$

Thus, the above implies that

$$\mathbb{E}\int_{\mathbb{R}^n}\left|\bar{V}(t,x)-\widetilde{V}(t,x)\right|^2dx\leq C\mathbb{E}\int_t^T\int_{\mathbb{R}^n}\left|\bar{V}(r,x)-\widetilde{V}(r,x)\right|^2dxdr.\tag{5.7}$$

Hence, by Gronwall's inequality, we obtain that

$$\bar{V}(t,x)=\widetilde{V}(t,x),\qquad(t,x)\in[0,T]\times\mathbb{R}^n,\ \text{a.s.}\tag{5.8}$$

Further, (5.5) implies that

$$\bar{q}(t,x)=\widetilde{q}(t,x),\qquad(t,x)\in[0,T]\times\mathbb{R}^n,\ \text{a.s.}\tag{5.9}$$

Then, by virtue of (5.4), we have

$$\bar{\zeta}(t,x)=\widetilde{\zeta}(t,x),\qquad(t,x)\in[0,T]\times\mathbb{R}^n,\ \text{a.s.}\tag{5.10}$$

which proves our conclusion. □

To establish the existence of an adapted weak solution, we define

$$\eta(\rho)=\begin{cases}0, & \rho\in(-\infty,0]\bigcup(2,\infty),\\ \rho, & \rho\in(0,1],\\ 2-\rho, & \rho\in(1,2],\end{cases}\tag{5.11}$$

and define

$$\psi(\rho)=\int_0^\rho\int_0^\tau\eta(r)drd\tau=\int_0^\rho(\rho-r)\eta(r)dr,\qquad\rho\in\mathbb{R}.\tag{5.12}$$

Thus, $\psi:\mathbb{R}\to\mathbb{R}$ is C^2, nondecreasing, and convex.

Now, for any $\varepsilon > 0$, we consider the following semilinear BSPDE:

$$V^\varepsilon(t,x) = \int_t^T \Big\{ \frac{1}{2}\nabla \cdot \big(\sigma\sigma^T V^\varepsilon_x\big) + \big\langle \bar{b}, V^\varepsilon_x \big\rangle + \bar{b}^0 V^\varepsilon + \nabla \cdot (\sigma q^\varepsilon) + \big\langle \bar{\sigma}^0, q^\varepsilon \big\rangle \tag{5.13}$$
$$+ \bar{g} - \psi\left(\frac{V^\varepsilon}{\varepsilon}\right)\Big\}dr - \int_t^T \langle\, q^\varepsilon, dW(r)\,\rangle, \qquad (t,x) \in [0,T]\times\mathbb{R}^n.$$

The unknown of semilinear BSPDE (5.13) is the pair $(V^\varepsilon, q^\varepsilon)$ of $\mathbb{F}$-adapted random fields. The following is a special case of a relevant result found in [26].

Theorem 5.2. *Let* (H3) *hold. Then, semilinear BSPDE* (5.13) *admits unique adapted classical solution* $(V^\varepsilon, q^\varepsilon)$. *Moreover, for any* $p > 1$ *and any compact set* $K \subseteq \mathbb{R}^n$,

$$\mathbb{E}\left[\sup_{t\in[0,T],\, x\in K} |\partial^\alpha V^\varepsilon(t,x)|^p\right] + \mathbb{E}\int_0^T \sup_{x\in K} |\partial^\alpha q^\varepsilon(t,x)|^p dt < \infty \quad \forall |\alpha| \le 2. \tag{5.14}$$

We hope that unique adapted classical solution $(V^\varepsilon, q^\varepsilon)$ of (4.20) will converge to $(\bar{V}, \bar{q})$ in some sense, where $(\bar{V}, \bar{q}, \bar{\zeta})$ is the adapted weak solution to our BSPDVI (4.20). Moreover, it is a hope that value function V of Problem (S) can be identified by $\bar{V}$ through (4.10). However, we note that, in the above estimate (5.14), the bound of the left-hand side depends not only on compact set K, but also depends on $\varepsilon > 0$, in general. Hence, we first would like to establish some estimates for $(V^\varepsilon, q^\varepsilon)$ (on the whole space $[0,T]\times\mathbb{R}^n$), which are uniform in $\varepsilon > 0$. To this end, we begin with several lemmas whose technical proofs will be given in the appendix.

Lemma 5.3. *Let* $\theta : \mathbb{R} \to [0,\infty)$ *be convex and piecewise smooth. Suppose that*

$$0 = \theta(0) = \min_{\rho\in\mathbb{R}} \theta(\rho) \tag{5.15}$$

and

$$[\theta'(\rho)]^2 \le C\theta(\rho)\theta''(\rho) \qquad \text{a.e. } \rho \in \mathbb{R} \tag{5.16}$$

for some constant $C > 0$. *Let* $(V^\varepsilon, q^\varepsilon)$ *be the adapted classical solution to BSPDE* (5.13). *Then*

$$\mathbb{E}\int_{\mathbb{R}^n} \theta\left(V^\varepsilon(t,x)\right)dx + \mathbb{E}\!\int_t^T\!\!\int_{\mathbb{R}^n} \Big\{\theta''(V^\varepsilon(r,x))\,\big|\,\sigma(r,x)^T V^\varepsilon_x(r,x) + q^\varepsilon(r,x)\big|^2 \tag{5.17}$$
$$+ \theta'(V^\varepsilon(r,x))\psi\left(\frac{V^\varepsilon(r,x)}{\varepsilon}\right)\Big\}dxdr \le C\mathbb{E}\!\int_t^T\!\!\int_{\mathbb{R}^n} |\theta'(V^\varepsilon(r,x))|$$
$$\left(|V^\varepsilon(r,x)| + |\bar{g}(r,x)|\right)dxdr, \qquad t \in [0,T].$$

The above lemma can be used to establish several interesting estimates for $(V^\varepsilon, q^\varepsilon)$.

Lemma 5.4. *Let* (H3) *hold and* $(V^\varepsilon, q^\varepsilon)$ *be the classical solution to BSPDE* (5.13). *Then there exists a constant* $C > 0$, *independent of* $m \ge 1$ *and* $\varepsilon > 0$ *such* *that*

$$\sup_{t\in[0,T]}\mathbb{E}\!\int_{\mathbb{R}^n}|V^\varepsilon(t,x)|^{2m}dx+\mathbb{E}\!\int_0^T\!\!\int_{\mathbb{R}^n}m^2|V^\varepsilon(r,x)|^{2m-2}\Big[|\sigma(r,x)^TV^\varepsilon_x(r,x)+q^\varepsilon(r,x)|^2 + mV^\varepsilon(r,x)\psi\left(\frac{V^\varepsilon(r,x)}{\varepsilon}\right)\Big]dxdr\le Ce^{Cm}\mathbb{E}\!\int_0^T\!\!\int_{\mathbb{R}^n}|\bar g(r,x)|^{2m}dxdr,\quad \forall\varepsilon>0,\ m\ge 1,\tag{5.18}$$

$$\sup_{t\in[0,T]}\mathbb{E}\int_{\mathbb{R}^n}\left(V^\varepsilon(t,x)^+\right)^2dx+\mathbb{E}\int_0^T\int_{\mathbb{R}^n}\Big\{\left|\sigma(r,x)^TV^\varepsilon_x(r,x)+q^\varepsilon(r,x)\right|^2I_{\{V^\varepsilon>0\}} + V^\varepsilon(r,x)^+\psi\left(\frac{V^\varepsilon(r,x)}{\varepsilon}\right)\Big\}dxdr\le C\mathbb{E}\int_0^T\int_{\mathbb{R}^n}|\bar g(r,x)|^2I_{\{V^\varepsilon>0\}}dxdr,\qquad\forall\varepsilon>0,\tag{5.19}$$

$$\sup_{t\in[0,T]}\mathbb{E}\int_{\mathbb{R}^n}V^\varepsilon(t,x)\psi\left(\frac{V^\varepsilon(t,x)}{\varepsilon}\right)dx + \mathbb{E}\int_0^T\int_{\mathbb{R}^n}\psi\left(\frac{V^\varepsilon(r,x)}{\varepsilon}\right)^2dxdr\le C\mathbb{E}\int_0^T\int_{\mathbb{R}^n}|\bar g(r,x)|^2dxdr.\tag{5.20}$$

Next, differentiating BSPDE (5.13) with respect to x_k, we get

$$\begin{aligned}V^\varepsilon_{x_k}(t,x)=\int_t^T\Big\{&\frac12\nabla\cdot\left(\sigma\sigma^T\left(V^\varepsilon_{x_k}\right)_x\right)+\left\langle\bar b,\left(V^\varepsilon_{x_k}\right)_x\right\rangle+\bar b^0V^\varepsilon_{x_k}+\nabla\cdot\left(\sigma q^\varepsilon_{x_k}\right)+\left\langle\bar\sigma^0,q^\varepsilon_{x_k}\right\rangle\\&-\psi'\left(\frac{V^\varepsilon}{\varepsilon}\right)\frac{V^\varepsilon_{x_k}}{\varepsilon}+\frac12\nabla\cdot\left(\left(\sigma\sigma^T\right)_{x_k}V^\varepsilon_x\right)+\left\langle\bar b_{x_k},V^\varepsilon_x\right\rangle+\bar b^0_{x_k}V^\varepsilon\\&+\nabla\cdot\left(\sigma_{x_k}q^\varepsilon\right)+\left\langle\bar\sigma^0_{x_k},q^\varepsilon\right\rangle+\bar g_{x_k}\Big\}dr-\int_t^T\left\langle q^\varepsilon_{x_k},dW(r)\right\rangle.\end{aligned}\tag{5.21}$$

We have the following result.

LEMMA 5.5. *Let* (H3) *hold and* $(V^\varepsilon,q^\varepsilon)$ *be the classical solution to BSPDE* (5.13). *Then there exists a constant* $C>0$, *independent of* $\varepsilon>0$ *such that*

$$\begin{aligned}&\mathbb{E}\!\int_{\mathbb{R}^n}|V^\varepsilon_x(t,x)|^2dx\\&\quad+\mathbb{E}\int_t^T\int_{\mathbb{R}^n}\Big\{\left|\sigma(r,x)^TV^\varepsilon_{xx}(r,x)+q^\varepsilon_x(r,x)^T\right|^2+\frac{|V^\varepsilon_x(r,x)|^2}{\varepsilon}\psi'\left(\frac{V^\varepsilon(r,x)}{\varepsilon}\right)\Big\}dxdr\\&\qquad\le C\mathbb{E}\int_t^T\int_{\mathbb{R}^n}\left(|\bar g(r,x)|^2+|\bar g_x(r,x)|^2\right)dxdr\qquad\forall\varepsilon>0,t\in[0,T].\end{aligned}\tag{5.22}$$

Note that, in [16] (see [17] also), to obtain an estimate similar to (5.22), a symmetric condition was assumed. Our proof above removes such a condition. Next, the result gives the monotonicity of the sequence $\{V^\varepsilon(\cdot\,,\cdot)\}_{\varepsilon>0}$ in $\varepsilon>0$.

LEMMA 5.6. *Let* (H3) *hold and* $0<\varepsilon<\delta$. *Let* $(V^\varepsilon,q^\varepsilon)$ *be the adapted classical solution of BSPDE* (5.13) *and* (V^δ,q^δ) *be the adapted classical solution of* (5.13) *with* ε *replaced by* δ. *Then,*

$$V^\varepsilon(t,x)\le V^\delta(t,x),\qquad(t,x)\in[0,T]\times\mathbb{R}^n,\ \text{a.s.}\tag{5.23}$$

Proof. We observe that, by the definition of $\psi(\cdot)$, one has

$$\psi\left(\frac{v}{\varepsilon}\right)\ge\psi\left(\frac{v}{\delta}\right),\qquad\forall 0<\varepsilon<\delta,\quad v\in\mathbb{R}.\tag{5.24}$$

Hence, by letting

$$V^{\varepsilon,\delta} = V^\varepsilon - V^\delta, \qquad q^{\varepsilon,\delta} = q^\varepsilon - q^\delta, \tag{5.25}$$

we have

$$\begin{aligned} V^{\varepsilon,\delta}(t,x) = \int_t^T \Big\{ &\frac{1}{2}\nabla \cdot \big(\sigma\sigma^T V_x^{\varepsilon,\delta}\big) + \big\langle \bar b, V_x^{\varepsilon,\delta}\big\rangle + \bar b^0 V^{\varepsilon,\delta} + \nabla\cdot\big(\sigma q^{\varepsilon,\delta}\big) + \big\langle \bar\sigma^0, q^{\varepsilon,\delta}\big\rangle \\ &- \Big[\psi\Big(\frac{V^\varepsilon}{\varepsilon}\Big) - \psi\Big(\frac{V^\delta}{\delta}\Big)\Big]\Big\} dr - \int_t^T \big\langle q^{\varepsilon,\delta}, dW(r)\big\rangle, \quad (t,x)\in[0,T]\times\mathbb{R}^n. \end{aligned} \tag{5.26}$$

Note that

$$\begin{aligned} -\Big[\psi\Big(\frac{V^\varepsilon}{\varepsilon}\Big) - \psi\Big(\frac{V^\delta}{\delta}\Big)\Big] &= -\Big[\psi\Big(\frac{V^\varepsilon}{\varepsilon}\Big) - \psi\Big(\frac{V^\delta}{\varepsilon}\Big) + \psi\Big(\frac{V^\delta}{\varepsilon}\Big) - \psi\Big(\frac{V^\delta}{\delta}\Big)\Big] \\ &= -\Big[\int_0^1 \psi'\Big(\frac{\lambda V^\varepsilon + (1-\lambda)V^\delta}{\varepsilon}\Big) d\lambda\Big] \frac{V^{\varepsilon,\delta}}{\varepsilon} \\ &\quad - \Big[\psi\Big(\tfrac{V^\delta}{\varepsilon}\Big) - \psi\Big(\tfrac{V^\delta}{\delta}\Big)\Big], \end{aligned} \tag{5.27}$$

with

$$-\Big[\psi\Big(\frac{V^\delta}{\varepsilon}\Big) - \psi\Big(\frac{V^\delta}{\delta}\Big)\Big] \le 0. \tag{5.28}$$

Hence, by a comparison theorem for linear BSPDEs (see [16]), we have

$$V^{\varepsilon,\delta}(t,x) \le 0, \qquad (t,x)\in[0,T]\times\mathbb{R}^n \quad \text{a.s.}, \tag{5.29}$$

which proves our conclusion. □

Having estimates (5.18)–(5.20), (5.22), and (5.23) for $(V^\varepsilon, q^\varepsilon)$, we are now ready to prove the following result.

THEOREM 5.7. *Let* (H3) *hold. Then, BSPDVI* (4.20) *admits an adapted weak solution* $(\bar V, \bar q, \bar\zeta)$.

Proof. First of all, by Lemma 5.6, we see that

$$\lim_{\varepsilon\to 0} V^\varepsilon(t,x) = \bar V(t,x) \qquad \forall (t,x)\in[0,T]\times\mathbb{R}^n \text{ a.s.} \tag{5.30}$$

for some $\mathbb{F}$-adapted random field $\bar V(\cdot\,,\cdot)$, with $(t,x)\mapsto \bar V(t,x)$ being upper semicontinuous a.s. Next, by taking the $2m$th root in both sides of (5.18) and then sending $m\to\infty$, we get

$$\operatorname*{esssup}_{(t,x,\omega)\in[0,T]\times\mathbb{R}^n\times\Omega} |V^\varepsilon(t,x,\omega)| \le C \operatorname*{esssup}_{(t,x,\omega)\in[0,T]\times\mathbb{R}^n\times\Omega} |\bar g(t,x,\omega)| \qquad \forall \varepsilon>0. \tag{5.31}$$

Hence, by taking $m=1$ in (5.18) and combining it with (5.22) and (5.31), we have

$$\begin{aligned} &\operatorname*{esssup}_{(t,x,\omega)\in[0,T]\times\mathbb{R}^n\times\Omega} |V^\varepsilon(t,x,\omega)|^2 + \sup_{t\in[0,T]} \mathbb{E}\int_{\mathbb{R}^n} |V_x^\varepsilon(t,x)|^2 dx \\ &\quad + \mathbb{E}\int_0^T\!\!\int_{\mathbb{R}^n} \Big\{ \big|\sigma(r,x)^T V_{xx}^\varepsilon(r,x) + q_x^\varepsilon(r,x)^T\big|^2 + |q^\varepsilon(r,x)|^2 \\ &\quad + \frac{V^\varepsilon(r,x)[V^\varepsilon(r,x)-\varepsilon] + |V_x^\varepsilon(r,x)|^2}{\varepsilon} I_{\{V^\varepsilon \ge 2\varepsilon\}}\Big\} dxdr \\ &\le C\Big[\operatorname*{esssup}_{(t,x,\omega)\in[0,T]\times\mathbb{R}^n\times\Omega} |\bar g(t,x,\omega)|^2 + \mathbb{E}\int_0^T\int_{\mathbb{R}^n} |\bar g_x(r,x)|^2 dxdr\Big] \qquad \forall\varepsilon>0. \end{aligned} \tag{5.32}$$

Next, by (5.32) and (5.20), together with the above, we know that with the $\bar{V}$ as in (5.30), and for some $\bar{q}$ and $\bar{\zeta}$, one has

$$\begin{cases} V^\varepsilon \to \bar{V} & \text{strongly in } L^2_{\mathbb{F}}\big(0,T;L^2(\mathbb{R}^n)\big), \\ V^\varepsilon_x \to \bar{V}_x & \text{weakly in } L^2_{\mathbb{F}}\big(0,T;L^2\left(\mathbb{R}^n;\mathbb{R}^{1\times n}\right)\big), \\ q^\varepsilon \to \bar{q} & \text{weakly in } L^2_{\mathbb{F}}\big(0,T;L^2\left(\mathbb{R}^n;\mathbb{R}^d\right)\big), \\ \psi\left(\dfrac{V^\varepsilon}{\varepsilon}\right) \to \bar{\zeta} & \text{weakly in } L^2_{\mathbb{F}}\big(0,T;L^2(\mathbb{R}^n)\big), \end{cases} \tag{5.33}$$

and

$$\mathbb{E}\int_0^T \int_{\mathbb{R}^n} \left\{[V^\varepsilon(r,x)-\varepsilon]^2 + |V^\varepsilon_x(r,x)|^2\right\} I_{(V^\varepsilon \geq 2\varepsilon)}dxdr \leq C\varepsilon. \tag{5.34}$$

This yields

$$\begin{cases} \left[(V^\varepsilon-\varepsilon)+|V^\varepsilon_x|\right] I_{\{V^\varepsilon\geq 2\varepsilon\}} \to 0 & \text{strongly in } L^2_{\mathbb{F}}\left(0,T;L^2\left(\mathbb{R}^n\right)\right), \\ \left[V^\varepsilon(t,x)+|V^\varepsilon_x(t,x)|\right] I_{\{V^\varepsilon\geq 2\varepsilon\}} \to 0 & \text{a.e. } (t,x)\in[0,T]\times\mathbb{R}^n \text{ a.s.} \end{cases} \tag{5.35}$$

Then it is necessary that

$$\bar{V}(t,x) \leq 0, \qquad (t,x)\in[0,T]\times\mathbb{R}^n, \quad \text{a.s.}, \tag{5.36}$$

and together with (5.30), we have

$$\big\{\bar{V}<0\big\} = \bigcup_{\varepsilon>0}\{V^\varepsilon<0\}. \tag{5.37}$$

Also, it is necessary that

$$\bar{\zeta}(t,x) \geq 0, \qquad \text{a.e. } (t,x)\in[0,T]\times\mathbb{R}^n, \text{ a.s.} \tag{5.38}$$

On the other hand, by applying the dominated convergence theorem to (5.19), we obtain

$$\begin{cases} (V^\varepsilon)^+\psi\left(\dfrac{V^\varepsilon}{\varepsilon}\right) = V^\varepsilon\psi\left(\dfrac{V^\varepsilon}{\varepsilon}\right) \to 0 & \text{strongly in } L^1_{\mathbb{F}}\big(0,T;L^1(\mathbb{R}^n)\big), \\ V^\varepsilon(t,x)\psi\left(\dfrac{V^\varepsilon(t,x)}{\varepsilon}\right) \to 0 & \text{a.e. } (t,x)\in[0,T]\times\mathbb{R}^n \text{ a.s.} \end{cases} \tag{5.39}$$

Hence, it is necessary that

$$\bar{V}(t,x)\bar{\zeta}(t,x) = 0, \qquad \text{a.e. } (t,x)\in[0,T]\times\mathbb{R}^n, \text{ a.s.} \tag{5.40}$$

Now, for any $\varphi(\cdot)\in C_0^\infty(\mathbb{R}^n)$, we have from (5.13) that

$$\begin{aligned} \int_{\mathbb{R}^n} V^\varepsilon(t,x)\varphi(x)dx = & \int_t^T\int_{\mathbb{R}^n}\bigg\{ -\left\langle \frac{1}{2}\sigma^T V^\varepsilon_x + q^\varepsilon, \sigma^T\varphi_x\right\rangle + \bigg[\big\langle \bar{b}, V^\varepsilon_x\big\rangle + \bar{b}^0 V^\varepsilon \\ & + \big\langle \bar{\sigma}^0, q^\varepsilon\big\rangle + \bar{g} - \psi\left(\frac{V^\varepsilon}{\varepsilon}\right)\bigg]\varphi\bigg\}dxdr \\ & - \int_t^T \left\langle \int_{\mathbb{R}^n} q^\varepsilon\varphi dx, dW(r)\right\rangle, \quad t\in[0,T]. \end{aligned} \tag{5.41}$$

Then, letting $\varepsilon \to 0$, along a sequence, we obtain

$$\begin{aligned}\int_{\mathbb{R}^n} \bar V(t,x)\varphi(x)dx = \int_t^T \int_{\mathbb{R}^n} \Big\{ &- \Big\langle \frac{1}{2}\sigma^T \bar V_x + \bar q, \sigma^T \varphi_x \Big\rangle \\ &+ \ \big[\langle \bar b, \bar V_x\rangle + \bar b^0 \bar V + \langle \bar\sigma^0, \bar q\rangle + \bar g - \bar\zeta\big]\, \varphi \Big\} dxdr \\ &- \int_t^T \Big\langle \int_{\mathbb{R}^n} \bar q \varphi dx, dW(r) \Big\rangle, \quad t \in [0,T].\end{aligned} \tag{5.42}$$

Hence, $(\bar V, \bar q, \bar\zeta)$ is the adapted weak solution of (4.20). □

From the above, we see that, for any $\varphi(\cdot) \in C_0^\infty(\mathbb{R}^n)$, map $t \mapsto \int_{\mathbb{R}^n} \bar V(t,x)\varphi(x)dx$ is continuous.

6. Identification of the value function. In this section, we are going to identify the weak adapted solution of BSPDVI as the value function $V(\cdot\,,\cdot)$ of Problem (S). Suppose $(V^\varepsilon(\cdot), q^\varepsilon(\cdot))$ is the adapted classical solution of (5.13). Let

$$\begin{cases} \widetilde V^\varepsilon(t,x) = (1+|x|^p)V^\varepsilon(t,x) + h(t,x), \\ \widetilde q^{\,\varepsilon}(t,x) = (1+|x|^p)q^\varepsilon(t,x) + \mu(t,x), \end{cases} \qquad (t,x) \in [0,T] \times \mathbb{R}^n. \tag{6.1}$$

Then

$$\begin{aligned}\widetilde V^\varepsilon(t,x) = h(T,x) &+ \int_t^T \Big\{ \frac{1}{2}\mathrm{tr}\, \Big[\sigma(r,x)\sigma(r,x)^T \widetilde V^\varepsilon_{xx}(r,x)\Big] + \Big\langle b(r,x), \widetilde V^\varepsilon_x(r,x)\Big\rangle \\ &+ \mathrm{tr}\,[\sigma(r,x)\widetilde q^{\,\varepsilon}_x(r,x)] - \psi\left(\frac{\widetilde V^\varepsilon(r,x) - h(r,x)}{\varepsilon}\right) + g(r,x)\Big\} dr \\ &- \int_t^T \langle\, \widetilde q^{\,\varepsilon}(r,x), dW(r)\,\rangle \\ \equiv h(T,x) &+ \int_t^T \widetilde q^{\,0,\varepsilon}(r,x)dr - \int_t^T \langle\, \widetilde q^{\,\varepsilon}(r,x), dW(r)\,\rangle, \qquad (t,x) \in [s,T] \times \mathbb{R}^n.\end{aligned} \tag{6.2}$$

Consequently, by Itô–Kunita's formula, we have

$$\begin{aligned}\widetilde V^\varepsilon(t, X(t;s,\xi)) = \widetilde V^\varepsilon(s,\xi) &+ \int_s^t \Big\{ \widetilde q^{\,0,\varepsilon}(r, X(r;s,\xi)) \\ &+ \ \Big\langle b(r,X(r;s,\xi)), \widetilde V^\varepsilon_x(r, X(r;s,\xi))\Big\rangle \\ &+ \frac{1}{2}\mathrm{tr}\, \Big[\sigma(r,X(r;s,\xi))\sigma(r,X(r;s,\xi))^T \widetilde V^\varepsilon_{xx}(r,X(r;s,\xi))\Big] \\ &+ \mathrm{tr}\,[\sigma(r,X(r;s,\xi))\widetilde q^{\,\varepsilon}_x(r,X(r;s,\xi))]\Big\} dr \\ &+ \int_s^t \Big\langle \widetilde q^{\,\varepsilon}(r,X(r;s,\xi)) + \sigma(r,X(r;s,\xi))^T \widetilde V^\varepsilon_x(r,X(r;s,\xi)), dW(r)\Big\rangle \\ = \widetilde V^\varepsilon(s,\xi) &+ \int_s^t \Big\{ \psi\left(\frac{\widetilde V^\varepsilon(r,X(r;s,\xi)) - h(r,X(r;s,\xi))}{\varepsilon}\right) \\ &\qquad - g(r,X(r;s,\xi))\Big\} dr \\ &+ \int_s^t \Big\langle \widetilde q^{\,\varepsilon}(r,X(r;s,\xi)) + \sigma(r,X(r;s,\xi))^T \widetilde V^\varepsilon_x(r,X(r;s,\xi)), dW(r)\Big\rangle.\end{aligned} \tag{6.3}$$

Thus, for any $\tau \in \mathcal{S}[s,T]$,

$$\begin{aligned} J_{s,\xi}(\tau) &= \mathbb{E}\left[\int_s^\tau g(r,X(r;s,\xi))dr + h(\tau,X(\tau;s,\xi))\big|\mathcal{F}_s\right] \\ &= \widetilde{V}^\varepsilon(s,\xi) + \mathbb{E}\left[\int_s^\tau \psi\left(\frac{\widetilde{V}^\varepsilon(r,X(r;s,\xi)) - h(r,X(r;s,\xi))}{\varepsilon}\right) dr\big|\mathcal{F}_s\right] \\ &\quad + \mathbb{E}\left[h(\tau,X(\tau;s,\xi)) - \widetilde{V}^\varepsilon(\tau,X(\tau,s,\xi))\big|\mathcal{F}_s\right]. \end{aligned} \tag{6.4}$$

By our discussion above, we know that under (H3),

$$\widetilde{V}^\varepsilon(t,x) \downarrow V^*(t,x), \qquad \forall (t,x) \in [0,T]\times\mathbb{R}^n, \text{ a.s.,} \tag{6.5}$$

with $(t,x) \mapsto V^*(t,x)$ being upper semicontinuous, and by (5.36),

$$V^*(s,\xi) = (1+|x|^p)\bar{V}(s,\xi) + h(s,\xi) \le h(s,\xi), \qquad \text{a.s., } \forall (s,\xi) \in \mathcal{D}^p. \tag{6.6}$$

Now, sending $\varepsilon \to 0$ in (6.4), we have

$$\begin{aligned} J_{s,\xi}(\tau) = V^*(s,\xi) &+ \mathbb{E}\left[\int_s^\tau \zeta^*(r,X(r;s,\xi))dr\big|\mathcal{F}_s\right] \\ &+ \mathbb{E}\left[h(\tau,X(\tau;s,\xi)) - V^*(\tau,X(\tau,s,\xi))\big|\mathcal{F}_s\right], \end{aligned} \tag{6.7}$$

with

$$\zeta^*(t,x) \in \beta\left(V^*(t,x) - h(t,x)\right), \qquad (t,x) \in [0,T]\times\mathbb{R}^n, \text{ a.s.} \tag{6.8}$$

Hence,

$$J_{s,\xi}(\tau) \ge V^*(s,\xi), \qquad (s,\xi) \in \mathcal{D}^p, \quad \tau \in \mathcal{S}[s,T], \tag{6.9}$$

which leads to

$$V(s,\xi) \ge V^*(s,\xi), \qquad (s,\xi) \in \mathcal{D}^p. \tag{6.10}$$

Next, let $(s,\xi) \in \mathcal{D}^p$ be fixed. We define

$$\tau^*(s,\xi) = \inf\left\{r \in [s,T] \;\big|\; V^*(r,X(r;s,\xi)) = h(r,X(r;s,\xi))\right\}. \tag{6.11}$$

Since $t \mapsto V^*(r,X(r;s,\xi)) - h(r,X(r;s,\xi))$ is $\mathbb{F}$-progressively measurable, by Début Theorem (see [6], [19]), $\tau^*(s,\xi) \in \mathcal{S}[s,T]$. Then, taking $\tau = \tau^*(s,\xi)$ in (6.7), we obtain

$$V(s,\xi) \le J_{s,\xi}(\tau^*(s,\xi)) = V^*(s,\xi). \tag{6.12}$$

Hence, combining the above with (6.10), one must have the following:

$$V^*(s,\xi) = V(s,\xi) \qquad \forall (s,\xi) \in \mathcal{D}^p. \tag{6.13}$$

That means $V^*(\cdot\,,\cdot)$ must be the value function $V(\cdot\,,\cdot)$ of Problem (S). Consequently, $(t,x) \mapsto V^*(t,x) = V(t,x)$ must be continuous itself. Moreover, the smallest optimal stopping time can be identified through (3.4).

7. Appendix. In this appendix, we collect some proofs.

Sketch proof of Theorem 3.1. (i) For any $\tau \in \mathcal{S}[s,T]$, we have

$$V(s,\xi) \le J_{s,\xi}(\tau) \equiv \mathbb{E}\left[\int_s^\tau g(r, X(r;s,\xi))dr + h(\tau, X(\tau))\big|\mathcal{F}_s\right]. \tag{7.1}$$

In particular, taking $\tau = s$, one obtains (3.1). Next, for any $\tau \in \mathcal{S}[s,T]$, take $\theta \in \mathcal{S}[\tau,T]$. One has

$$V(s,\xi) \le J_{s,\xi}(\theta) = \mathbb{E}\left[\int_s^\tau g(r, X(r;s,\xi))dt + J_{\tau,X(\tau;s,\xi)}(\theta)\big|\mathcal{F}_s\right]. \tag{7.2}$$

Taking infimum with respect to $\theta \in \mathcal{S}[\tau,T]$ yields

$$V(s,\xi) \le \mathbb{E}\left[\int_s^\tau g(r, X(r;s,\xi))dr + V(\tau, X(\tau;s,\xi))\big|\mathcal{F}_s\right]. \tag{7.3}$$

Hence, (3.2) holds.

(ii) Suppose $\bar\theta \in \mathcal{S}[\theta,T]$ is optimal for initial point $(s,\xi) \in \mathcal{D}^p$. Then

$$\begin{aligned} V(s,\xi) = J_{s,\xi}(\bar\theta) &\ge \mathbb{E}\left[\int_s^{\bar\theta} g(r, X(r;s,\xi))dr + V(\bar\theta, X(\bar\theta;s,\xi))\big|\mathcal{F}_s\right] \\ &\ge \inf_{\tau\in\mathcal{S}[s,T]} \mathbb{E}\left[\int_s^\tau g(r, X(r;s,\xi))dr + V(\tau, X(\tau;s,\xi))\big|\mathcal{F}_s\right] \ge V(s,\xi). \end{aligned} \tag{7.4}$$

Hence, the equalities in the above have to hold, which implies

$$\mathbb{E}\left[V\left(\bar\theta, X\left(\bar\theta;s,\xi\right)\right)\big|\mathcal{F}_s\right] = \mathbb{E}\left[h\left(\bar\theta, X\left(\bar\theta;s,\xi\right)\right)\big|\mathcal{F}_s\right]. \tag{7.5}$$

Combining the fact

$$V\left(\bar\theta, X\left(\bar\theta;s,\xi\right)\right) \le h\left(\bar\theta, X\left(\bar\theta;s,\xi\right)\right) \qquad \text{a.s.},$$

we obtain (3.3). Next, for (3.5), if there exists a $\Omega_0 \subseteq \{V(s,\xi) < h(s,\xi)\}$, with $\mathbb{P}(\Omega_0) > 0$ such that

$$\bar\tau(s,\xi) = s \qquad \text{on } \ \Omega_0, \tag{7.6}$$

then, trivially,

$$V(s,\xi(\omega)) = h(s,\xi(\omega)), \qquad \omega \in \Omega_0, \tag{7.7}$$

which contradicts the choice of Ω_0. Conversely, if $\Omega_0 \subseteq \{\bar\tau(s,\xi) > s\}$, with $\mathbb{P}(\Omega_0) > 0$ such that (7.7) holds, then (7.6) has to be true (by definition of $\bar\tau(s,\xi)$), a contradiction to the choice of Ω_0. Hence, (3.5) holds.

We now show (3.6). To this end, let $(s,\xi) \in \mathcal{D}^p$. Define $\bar\tau(s,\xi)$ by (3.4), and suppose $\mathbb{P}\{s < \bar\tau(s,\xi)\} > 0$. The case $\theta = \bar\tau(s,\xi)$ is trivial. Thus, we fix $\theta \in \mathcal{S}[s,\bar\tau(s,\xi))$, and let $\tau \in \mathcal{S}[\theta,\bar\tau(s,\xi)]$. From (3.3), we know that any $\mu \in \mathcal{S}[\theta,T]$, with $\mathbb{P}\{\mu < \tau\} > 0$ is not optimal for initial point $(\theta, X(\theta;s,\xi))$. Hence,

$$\begin{aligned} V(\theta, X(\theta;s,\xi)) &= \inf_{\mu\in\mathcal{S}[\tau,T]} \mathbb{E}\left[\int_\theta^\tau g(r, X(r;s,\xi))dr + J_{\tau,X(\tau;s,\xi)}(\mu)\big|\mathcal{F}_\theta\right] \\ &= \mathbb{E}\left[\int_\theta^\tau g(r, X(r;s,\xi))dr + V(\tau, X(\tau;s,\xi))\big|\mathcal{F}_\theta\right], \end{aligned} \tag{7.8}$$

proving (3.6).

Finally, by taking $\theta = s$ and $\tau = \bar\tau(s,\xi)$ in (3.6), we see that

$$V(s,\xi) = \mathbb{E}\left[\int_s^{\bar\tau(s,x)} g(s,X(s;s,\xi))ds + h(\bar\tau(s,\xi),X(\bar\tau(s,x);s,\xi))\big|\mathcal{F}_s\right] = J_{s,\xi}(\bar\tau(s,\xi)), \tag{7.9}$$

which means that $\bar\tau(s,\xi)$ is an optimal stopping time of Problem (S) for the initial point (s,ξ), and it must be the smallest one. □

Proof of Lemma 5.3. By (5.16), we may assume that

$$\frac{\theta'(\rho)}{\theta''(\rho)} = \frac{\theta'(\rho)}{\theta''(\rho)} I_{\{\theta''(\rho)>0\}} \qquad \forall \rho \in \mathbb{R}, \tag{7.10}$$

since $\theta : \mathbb{R} \to [0,\infty)$ is convex and piecewise smooth. Applying Itô's formula to $\theta(V^\varepsilon)$, we have

$$\begin{aligned}
&-\mathbb{E}\textstyle\int_{\mathbb{R}^n} \theta\left(V^\varepsilon(t,x)\right)dx\\
&\quad= -\mathbb{E}\int_t^T\int_{\mathbb{R}^n}\Big\{\theta'(V^\varepsilon)\Big[\frac{1}{2}\nabla\cdot\left(\sigma\sigma^T V_x^\varepsilon\right) + \left\langle \bar b, V_x^\varepsilon\right\rangle + \bar b^0 V^\varepsilon + \nabla\cdot(\sigma q^\varepsilon) + \left\langle \bar\sigma^0, q^\varepsilon\right\rangle\\
&\qquad + \bar g - \psi\left(\tfrac{V^\varepsilon}{\varepsilon}\right)\Big] - \tfrac{1}{2}\theta''(V^\varepsilon)|q^\varepsilon|^2\Big\}dxdr\\
&\quad= \frac{1}{2}\mathbb{E}\int_t^T\int_{\mathbb{R}^n}\Big\{\theta''(V^\varepsilon)\left|\sigma^T V_x^\varepsilon\right|^2 + 2\left\langle q^\varepsilon, \theta''(V^\varepsilon)\sigma^T V_x^\varepsilon - \theta'(V^\varepsilon)\bar\sigma^0\right\rangle + \theta''(V^\varepsilon)|q^\varepsilon|^2\\
&\qquad + 2\left(\nabla\cdot\bar b\right)\theta(V^\varepsilon) + 2\theta'(V^\varepsilon)\left[\psi\left(\frac{V^\varepsilon}{\varepsilon}\right) - \bar b^0 V^\varepsilon - \bar g\right]\Big\}dxdr\\
&\quad= \frac{1}{2}\mathbb{E}\int_t^T\int_{\mathbb{R}^n}\Big\{\theta''(V^\varepsilon)\left[\left|\sigma^T V_x^\varepsilon\right|^2 + 2\left\langle q^\varepsilon, \sigma^T V_x^\varepsilon - \frac{\theta'(V^\varepsilon)}{\theta''(V^\varepsilon)}\bar\sigma^0\right\rangle + |q^\varepsilon|^2\right]\\
&\qquad + 2\left(\nabla\cdot\bar b\right)\theta(V^\varepsilon) + 2\theta'(V^\varepsilon)\left[\psi\left(\frac{V^\varepsilon}{\varepsilon}\right) - \bar b^0 V^\varepsilon - \bar g\right]\Big\}dxdr\\
&\quad= \frac{1}{2}\mathbb{E}\int_t^T\int_{\mathbb{R}^n}\Big\{\theta''(V^\varepsilon)\left|\sigma^T V_x^\varepsilon + q^\varepsilon - \frac{\theta'(V^\varepsilon)}{\theta''(V^\varepsilon)}\bar\sigma^0\right|^2 + 2\left[\left(\nabla\cdot\bar b\right) - \nabla\cdot\left(\sigma\bar\sigma^0\right)\right]\theta(V^\varepsilon)\\
&\qquad - \frac{|\theta'(V^\varepsilon)|^2}{\theta''(V^\varepsilon)}\left|\bar\sigma^0\right|^2 + \theta'(V^\varepsilon)\left[\psi\left(\frac{V^\varepsilon}{\varepsilon}\right) - \bar b^0 V^\varepsilon - \bar g\right]\Big\}dxdr\\
&\quad\geq \frac{1}{2}\mathbb{E}\int_t^T\int_{\mathbb{R}^n}\Big\{\frac{\theta''(V^\varepsilon)}{2}\left|\sigma^T V_x^\varepsilon + q^\varepsilon\right|^2 + 2\left[\left(\nabla\cdot\bar b\right) - \nabla\cdot\left(\sigma\bar\sigma^0\right)\right]\theta(V^\varepsilon)\\
&\qquad - \frac{2|\theta'(V^\varepsilon)|^2}{\theta''(V^\varepsilon)}\left|\bar\sigma^0\right|^2 + \theta'(V^\varepsilon)\left[\psi\left(\frac{V^\varepsilon}{\varepsilon}\right) - \bar b^0 V^\varepsilon - \bar g\right]\Big\}dxdr\\
&\quad\geq \frac{1}{2}\mathbb{E}\int_t^T\int_{\mathbb{R}^n}\Big\{\theta''(V^\varepsilon)\left|\sigma^T V_x^\varepsilon + q^\varepsilon\right|^2 + \theta'(V^\varepsilon)\psi\left(\frac{V^\varepsilon}{\varepsilon}\right)\\
&\qquad - C\theta(V^\varepsilon) - C|\theta'(V^\varepsilon)|\left(|V^\varepsilon| + |\bar g|\right)\Big\}dxdr.
\end{aligned} \tag{7.11}$$

In the above, we have used the fact that $|a-b| \geq \frac{1}{2}|a| - |b|$. Note that, under our conditions,

$$\theta'(\rho)\psi\left(\frac{\rho}{\varepsilon}\right) \geq 0 \qquad \text{a.e. } \rho \in \mathbb{R}. \tag{7.12}$$

Hence, by Gronwall's inequality, we obtain (5.17). □

Proof of Lemma 5.4. For any $m \geq 1$, taking

$$\theta(\rho) = |\rho|^{2m}, \qquad \rho \in \mathbb{R}^n. \tag{7.13}$$

Then (5.15) and (5.16) hold. Hence, by Lemma 5.3, we obtain

$$\begin{aligned} &\mathbb{E}\int_{\mathbb{R}^n} V^\varepsilon(t,x)^{2m}dx + \mathbb{E}\int_t^T\int_{\mathbb{R}^n} \Big\{2m(2m-1)(V^\varepsilon)^{2m-2}\,\big|\,\sigma^T V_x^\varepsilon + q^\varepsilon\big|^2 \\ &\quad + 2m(V^\varepsilon)^{2m-1}\psi\left(\frac{V^\varepsilon}{\varepsilon}\right)\Big\}dxdr \leq C\mathbb{E}\int_t^T\int_{\mathbb{R}^n} 2m|V^\varepsilon|^{2m-1}|\,(|V^\varepsilon| + |\bar{g}|)\,dxdr \\ &\quad \leq C\mathbb{E}\int_t^T\int_{\mathbb{R}^n} \left\{(4m-1)(V^\varepsilon)^{2m} + |\bar{g}|^{2m}\right\}dxdr. \end{aligned} \tag{7.14}$$

Then, by Gronwall's inequality, one has

$$\begin{aligned} &\mathbb{E}\int_{\mathbb{R}^n} V^\varepsilon(t,x)^{2m}dx + \mathbb{E}\int_t^T\int_{\mathbb{R}^n} \Big\{2m(2m-1)(V^\varepsilon)^{2m-2}\,\big|\,\sigma^T V_x^\varepsilon + q^\varepsilon\big|^2 \\ &\qquad + 2m(V^\varepsilon)^{2m-1}\psi\left(\frac{V^\varepsilon}{\varepsilon}\right)\Big\}dxdr \leq Ce^{Cm}\mathbb{E}\int_t^T\int_{\mathbb{R}^n} \bar{g}^{2m}dxdr, \end{aligned} \tag{7.15}$$

which implies (5.18).

Next, by taking

$$\theta(\rho) = \left(\rho^+\right)^2, \qquad \rho \in \mathbb{R}, \tag{7.16}$$

we have

$$\theta'(\rho) = 2\rho^+, \quad \theta''(\rho) = 2I_{\{\rho>0\}}, \tag{7.17}$$

which leads to

$$\theta'(\rho)^2 = 4\left(\rho^+\right)^2 = 2\theta(\rho)\theta''(\rho). \tag{7.18}$$

Thus, (5.15) and (5.16) hold. Hence, by Lemma 5.3, we have

$$\begin{aligned} &\mathbb{E}\int_{\mathbb{R}^n} \left(V^\varepsilon(t,x)^+\right)^2 dx + \mathbb{E}\int_t^T\int_{\mathbb{R}^n} \left\{2I_{\{V^\varepsilon>0\}}\,\big|\,\sigma^T V_x^\varepsilon + q^\varepsilon\big|^2 + 2(V^\varepsilon)^+\psi\left(\frac{V^\varepsilon}{\varepsilon}\right)\right\}dxdr \\ &\leq C\mathbb{E}\int_t^T\int_{\mathbb{R}^n} 2(V^\varepsilon)^+|\,[|V^\varepsilon| + |\bar{g}|]\,dxdr \leq C\mathbb{E}\int_t^T\int_{\mathbb{R}^n} \left[\left((V^\varepsilon)^+\right)^2 + |\bar{g}|^2 I_{\{V^\varepsilon>0\}}\right]dxdr. \end{aligned} \tag{7.19}$$

It follows from Gronwall's inequality that

$$\begin{aligned} &\mathbb{E}\int_{\mathbb{R}^n} \left(V^\varepsilon(t,x)^+\right)^2 dx + \mathbb{E}\int_0^T\int_{\mathbb{R}^n} \left\{\big|\,\sigma^T V_x^\varepsilon + q^\varepsilon\big|^2 I_{\{V^\varepsilon>0\}} + (V^\varepsilon)^+\psi\left(\frac{V^\varepsilon}{\varepsilon}\right)\right\}dxdr \\ &\qquad \leq C\mathbb{E}\int_0^T\int_{\mathbb{R}^n} |\bar{g}(r,x)|^2 I_{\{V^\varepsilon>0\}}dxdr, \qquad \forall \varepsilon > 0, \end{aligned} \tag{7.20}$$

with $C > 0$ independent of $\varepsilon > 0$, which leads to (5.19).

Finally, we take

$$\theta(\rho)=\rho\psi\left(\frac{\rho}{\varepsilon}\right)=\frac{\rho^2}{\varepsilon}\int_0^{\rho/\varepsilon}\eta(r)dr-\rho\int_0^{\rho/\varepsilon}r\eta(r)dr.$$

Then

$$\begin{aligned}0\le\theta'(\rho)&=\psi\left(\frac{\rho}{\varepsilon}\right)+\frac{\rho}{\varepsilon}\psi'\left(\frac{\rho}{\varepsilon}\right)\\&=\frac{\rho}{\varepsilon}\int_0^{\rho/\varepsilon}\eta(r)dr-\int_0^{\rho/\varepsilon}r\eta(r)dr+\frac{\rho}{\varepsilon}\int_0^{\rho/\varepsilon}\eta(r)dr\\&=\frac{2\rho}{\varepsilon}\int_0^{\rho/\varepsilon}\eta(r)dr-\int_0^{\rho/\varepsilon}r\eta(r)dr,\end{aligned}\tag{7.21}$$

and

$$0\le\theta''(\rho)=\frac{2}{\varepsilon}\psi'\left(\frac{\rho}{\varepsilon}\right)+\frac{\rho}{\varepsilon^2}\psi''\left(\frac{\rho}{\varepsilon}\right)=\frac{2}{\varepsilon}\int_0^{\rho/\varepsilon}\eta(r)dr+\frac{\rho}{\varepsilon^2}\eta\left(\frac{\rho}{\varepsilon}\right).\tag{7.22}$$

Note that

$$\theta'(\rho)^2-C\theta(\rho)\theta''(\rho)=\left[\psi\left(\frac{\rho}{\varepsilon}\right)+\frac{\rho}{\varepsilon}\psi'\left(\frac{\rho}{\varepsilon}\right)\right]^2-C\psi\left(\frac{\rho}{\varepsilon}\right)\left[\frac{2\rho}{\varepsilon}\psi'\left(\frac{\rho}{\varepsilon}\right)+\frac{\rho^2}{\varepsilon^2}\psi''\left(\frac{\rho}{\varepsilon}\right)\right].\tag{7.23}$$

Now, for $\rho\in(-\infty,0]$, we have

$$\theta'(\rho)^2-C\theta(\rho)\theta''(\rho)=0\tag{7.24}$$

for any $C>0$. For $\rho\in(0,\varepsilon]$,

$$\theta'(\rho)^2-C\theta(\rho)\theta''(\rho)=\left[\frac{\rho^3}{6\varepsilon^3}+\frac{\rho}{\varepsilon}\frac{\rho^2}{2\varepsilon^2}\right]^2-C\frac{\rho^3}{6\varepsilon^3}\left[\frac{2\rho}{\varepsilon}\frac{\rho^2}{2\varepsilon^2}+\frac{\rho^3}{\varepsilon^3}\right]=\frac{\rho^6}{9\varepsilon^6}(4-3C)\le0,\tag{7.25}$$

provided $C\ge\frac{4}{3}$. For $\rho\in(\varepsilon,2\varepsilon]$, since both $\psi(\cdot)$ and $\psi'(\cdot)$ are nondecreasing,

$$\begin{aligned}\theta'(\rho)^2-C\theta(\rho)\theta''(\rho)&=\left[\psi\left(\frac{\rho}{\varepsilon}\right)+\frac{\rho}{\varepsilon}\psi'\left(\frac{\rho}{\varepsilon}\right)\right]^2-C\psi\left(\frac{\rho}{\varepsilon}\right)\left[\frac{2\rho}{\varepsilon}\psi'\left(\frac{\rho}{\varepsilon}\right)+\frac{\rho^2}{\varepsilon^2}\psi''\left(\frac{\rho}{\varepsilon}\right)\right]\\&\le\left[\psi(2)+2\psi'(2)\right]^2-2C\psi(1)\psi'(1)=9-\frac{C}{6}\le0,\end{aligned}\tag{7.26}$$

provided $C\ge54$, and for $\rho\in[2\varepsilon,\infty)$,

$$\begin{aligned}\theta'(\rho)^2-C\theta(\rho)\theta''(\rho)&=\left[\frac{2\rho}{\varepsilon}-1\right]^2-C\left(\frac{\rho}{\varepsilon}-1\right)\frac{2\rho}{\varepsilon}\le\frac{4\rho^2}{\varepsilon^2}-\frac{4\rho}{\varepsilon}+1-2C\frac{\rho^2}{\varepsilon^2}+2C\frac{\rho}{\varepsilon}\\&=-(C-2)\frac{2\rho}{\varepsilon}\left(\frac{\rho}{\varepsilon}-1\right)+1\le-4(C-2)+1=-4C+9\le0,\end{aligned}\tag{7.27}$$

provided $C\ge\frac{9}{4}$. Further, we claim that

$$\frac{\rho}{\varepsilon}\psi'\left(\frac{\rho}{\varepsilon}\right)\le3\psi\left(\frac{\rho}{\varepsilon}\right)+1,\qquad\rho\in\mathbb{R}.\tag{7.28}$$

In fact, the above holds for $\rho \le 0$. Now, for $\rho \in (0, \varepsilon]$,

$$\frac{\rho}{\varepsilon}\psi'\left(\frac{\rho}{\varepsilon}\right) = \frac{\rho^3}{2\varepsilon^3} = 3\psi\left(\frac{\rho}{\varepsilon}\right), \tag{7.29}$$

and for $\rho \in (\varepsilon, \infty)$,

$$\frac{\rho}{\varepsilon}\psi'\left(\frac{\rho}{\varepsilon}\right) = \frac{\rho}{\varepsilon}\left[1 - \frac{(2-\rho)^2}{2} I_{\{\rho \le 2\varepsilon\}}\right] \le \frac{\rho}{\varepsilon} = \psi\left(\frac{\rho}{\varepsilon}\right) + 1. \tag{7.30}$$

Hence, by Lemma 5.3,

$$\begin{aligned} &\mathbb{E}\int_{\mathbb{R}^n} V^\varepsilon(t,x)\psi\left(\frac{V^\varepsilon(t,x)}{\varepsilon}\right)dx + \mathbb{E}\int_t^T\int_{\mathbb{R}^n} \\ &\left\{\left[\frac{2}{\varepsilon}\psi'\left(\frac{V^\varepsilon}{\varepsilon}\right) + \frac{V^\varepsilon}{\varepsilon^2}\psi\left(\frac{V^\varepsilon}{\varepsilon}\right)\right]\left|\,\sigma^T V_x^\varepsilon + q^\varepsilon\right|^2\right. \\ &\qquad \left. + \left[\psi\left(\frac{V^\varepsilon}{\varepsilon}\right) + \frac{V^\varepsilon}{\varepsilon}\psi'\left(\frac{V^\varepsilon}{\varepsilon}\right)\right]\psi\left(\frac{V^\varepsilon}{\varepsilon}\right)\right\}dxdr \\ &\le C\mathbb{E}\int_t^T\int_{\mathbb{R}^n}\left[\psi\left(\frac{V^\varepsilon}{\varepsilon}\right) + \frac{V^\varepsilon}{\varepsilon}\psi'\left(\frac{V^\varepsilon}{\varepsilon}\right)\right](|V^\varepsilon| + |\bar g|)\,dxdr \\ &\le C\mathbb{E}\int_t^T\int_{\mathbb{R}^n}\psi\left(\frac{V^\varepsilon}{\varepsilon}\right)(V^\varepsilon + |\bar g|)\,dxdr, \qquad t \in [0,T]. \end{aligned} \tag{7.31}$$

Then, by Gronwall's inequality, together with Cauchy–Schwarz inequality, we obtain

$$\begin{aligned} \mathbb{E}\int_{\mathbb{R}^n} V^\varepsilon(t,x)\psi\left(\frac{V^\varepsilon(t,x)}{\varepsilon}\right)dx + \mathbb{E}\int_t^T\int_{\mathbb{R}^n}\psi\left(\frac{V^\varepsilon}{\varepsilon}\right)^2 dxdr& \\ \le C\mathbb{E}\int_t^T\int_{\mathbb{R}^n}|\bar g|^2 dxdr, \quad t \in [0,T]&. \end{aligned} \tag{7.32}$$

This leads to (5.20). □

Proof of Lemma 5.5. Applying Itô's formula to $|V_{x_k}^\varepsilon(t,x)|^2$ yields

$$\begin{aligned} &-\mathbb{E}\int_{\mathbb{R}^n}|V_{x_k}^\varepsilon(t,x)|^2 dx \\ &= -\,\mathbb{E}\int_t^T\int_{\mathbb{R}^n}\left\{2V_{x_k}^\varepsilon\left[\frac{1}{2}\nabla\cdot\left(\sigma\sigma^T\left(V_{x_k}^\varepsilon\right)_x\right) + \left\langle \bar b, \left(V_{x_k}^\varepsilon\right)_x\right\rangle + \bar b^0 V_{x_k}^\varepsilon + \nabla\cdot\left(\sigma q_{x_k}^\varepsilon\right)\right.\right. \\ &\quad + \left\langle \bar\sigma^0, q_{x_k}^\varepsilon\right\rangle - \psi'\left(\frac{V^\varepsilon}{\varepsilon}\right)\frac{V_{x_k}^\varepsilon}{\varepsilon} + \frac{1}{2}\nabla\cdot\left(\left(\sigma\sigma^T\right)_{x_k} V_x^\varepsilon\right) + \left\langle \bar b_{x_k}, V_x^\varepsilon\right\rangle + \bar b_{x_k}^0 V^\varepsilon \\ &\quad \left.\left. + \nabla\cdot(\sigma_{x_k} q^\varepsilon) + \left\langle \bar\sigma_{x_k}^0, q^\varepsilon\right\rangle + \bar g_{x_k}\right] - |q_{x_k}^\varepsilon|^2\right\}dxdr \\ &= \mathbb{E}\int_t^T\int_{\mathbb{R}^n}\left\{\left|\sigma^T\left(V_{x_k}^\varepsilon\right)_x\right|^2 + 2\left\langle q_{x_k}^\varepsilon, \sigma^T\left(V_{x_k}^\varepsilon\right)_x - V_{x_k}^\varepsilon\bar\sigma^0\right\rangle\right. \\ &\quad + |q_{x_k}^\varepsilon|^2 + 2\left\langle q^\varepsilon, \sigma_{x_k}^T\left(V_{x_k}^\varepsilon\right)_x - V_{x_k}^\varepsilon\bar\sigma_{x_k}^0\right\rangle + \left[(\nabla\cdot\bar b) - 2\bar b^0\right]\left(V_{x_k}^\varepsilon\right)^2 - 2V_{x_k}^\varepsilon\bar g_{x_k} \\ &\quad + 2\psi'\left(\frac{V^\varepsilon}{\varepsilon}\right)\frac{\left(V_{x_k}^\varepsilon\right)^2}{\varepsilon} + \left\langle (V_x^\varepsilon)_{x_k}, \left(\sigma\sigma^T\right)_{x_k} V_x^\varepsilon\right\rangle \\ &\quad \left. - 2V_{x_k}^\varepsilon\left\langle \bar b_{x_k}, V_x^\varepsilon\right\rangle + \bar b_{x_k x_k}^0 (V^\varepsilon)^2\right\}dxdr. \end{aligned} \tag{7.33}$$

Note that (recalling $\sigma = (\sigma_1, \ldots, \sigma_d)$, with each σ_i taking values in $\mathbb{R}^n$)

$$\begin{aligned}\nabla \cdot \left[V^\varepsilon_{x_k} \sigma_{x_k}\right] &= \big(\nabla \cdot [V^\varepsilon_{x_k}(\sigma_1)_{x_k}], \ldots, \nabla \cdot [V^\varepsilon_{x_k}(\sigma_d)_{x_k}]\big)^T \\ &= \big(\big\langle \big(V^\varepsilon_{x_k}\big)_x, (\sigma_1)_{x_k} \big\rangle + V^\varepsilon_{x_k} \nabla \cdot [(\sigma_1)_{x_k}], \ldots, \\ &\qquad \big\langle \big(V^\varepsilon_{x_k}\big)_x, (\sigma_d)_{x_k} \big\rangle + V^\varepsilon_{x_k} \nabla \cdot [(\sigma_d)_{x_k}]\big)^T \\ &= \sigma^T_{x_k} \left(V^\varepsilon_{x_k}\right)_x + V^\varepsilon_{x_k} \nabla \cdot (\sigma_{x_k}).\end{aligned} \tag{7.34}$$

Hence (recall that q^ε_x takes values in $\mathbb{R}^{d\times n}$),

$$\begin{aligned}\mathbb{E}\int_t^T \int_{\mathbb{R}^n} \left\langle q^\varepsilon, \sigma^T_{x_k} \left(V^\varepsilon_{x_k}\right)_x \right\rangle dxdr &= \mathbb{E}\int_t^T \int_{\mathbb{R}^n} \left\langle q^\varepsilon, \nabla \cdot \left[V^\varepsilon_{x_k}\sigma_{x_k}\right] - V^\varepsilon_{x_k} \nabla \cdot \sigma_{x_k} \right\rangle dxdr \\ &= -\mathbb{E}\int_t^T \int_{\mathbb{R}^n} V^\varepsilon_{x_k} \left\{ \operatorname{tr}\left(\sigma_{x_k} q^\varepsilon_x\right) + \langle\, q^\varepsilon, \nabla \cdot \sigma_{x_k} \,\rangle \right\} dxdr.\end{aligned} \tag{7.35}$$

On the other hand,

$$\begin{aligned}\left[\left\langle \left(\sigma\sigma^T\right)_{x_k} V^\varepsilon_x, V^\varepsilon_x \right\rangle\right]_{x_k} &= \left\langle \left(\sigma\sigma^T\right)_{x_kx_k} V^\varepsilon_x, V^\varepsilon_x \right\rangle + \left\langle \left(\sigma\sigma^T\right)_{x_k} (V^\varepsilon_x)_{x_k}, V^\varepsilon_x \right\rangle \\ &\quad + \left\langle \left(\sigma\sigma^T\right)_{x_k} V^\varepsilon_x, (V^\varepsilon_x)_{x_k} \right\rangle \\ &= \left\langle \left(\sigma\sigma^T\right)_{x_kx_k} V^\varepsilon_x, V^\varepsilon_x \right\rangle + 2\left\langle \left(\sigma\sigma^T\right)_{x_k} V^\varepsilon_x, (V^\varepsilon_x)_{x_k} \right\rangle,\end{aligned} \tag{7.36}$$

which implies that

$$\mathbb{E}\int_t^T \int_{\mathbb{R}^n} \left\langle (V^\varepsilon_x)_{x_k}, \left(\sigma\sigma^T\right)_{x_k} V^\varepsilon_x \right\rangle dxdr = -\frac{1}{2}\mathbb{E}\int_t^T \int_{\mathbb{R}^n} \left\langle \left(\sigma\sigma^T\right)_{x_kx_k} V^\varepsilon_x, V^\varepsilon_x \right\rangle dxdr. \tag{7.37}$$

Thus, (7.33) can be written as

$$\begin{aligned}&-\mathbb{E}\int_{\mathbb{R}^n} |V^\varepsilon_{x_k}(t,x)|^2 dx \\ &= \mathbb{E}\int_t^T \int_{\mathbb{R}^n} \Big\{ \left|\sigma^T (V^\varepsilon_{x_k})_x\right|^2 + 2\left\langle q^\varepsilon_{x_k}, \sigma^T \left(V^\varepsilon_{x_k}\right)_x - V^\varepsilon_{x_k}\bar\sigma^0 \right\rangle + |q^\varepsilon_{x_k}|^2 - 2V^\varepsilon_{x_k} \operatorname{tr}\left(\sigma_{x_k} q^\varepsilon_x\right) \\ &\quad + 2\left\langle q^\varepsilon, \bar\sigma^0_{x_k} - \nabla \cdot \sigma_{x_k} \right\rangle V^\varepsilon_{x_k} + \left[(\nabla \cdot \bar b) - 2\bar b^0\right] \left(V^\varepsilon_{x_k}\right)^2 - 2V^\varepsilon_{x_k}\bar g_{x_k} + 2\psi'\left(\frac{V^\varepsilon}{\varepsilon}\right) \frac{\left(V^\varepsilon_{x_k}\right)^2}{\varepsilon} \\ &\quad - \frac{1}{2}\left\langle \left(\sigma\sigma^T\right)_{x_kx_k} V^\varepsilon_x, V^\varepsilon_x \right\rangle - 2V^\varepsilon_{x_k} \left\langle \bar b_{x_k}, V^\varepsilon_x \right\rangle + \bar b^0_{x_kx_k} \left(V^\varepsilon\right)^2 \Big\} dxdr.\end{aligned} \tag{7.38}$$

In what follows, we let

$$\langle\, A, B \,\rangle = \operatorname{tr}\left(AB^T\right) = \sum_{k=1}^n a_k^T b_k, \qquad \forall A \equiv \begin{pmatrix} a_1^T \\ \vdots \\ a_n^T \end{pmatrix},\ B \equiv \begin{pmatrix} b_1^T \\ \vdots \\ b_n^T \end{pmatrix} \in \mathbb{R}^{n\times m}.$$

Then one has

$$|A|^2 = \operatorname{tr}\left(A^T A\right) = \sum_{i=1}^n \sum_{j=1}^m |a_{ij}|^2 \qquad \forall A \equiv (a_{ij}) \in \mathbb{R}^{n\times m}.$$

Now, summing (7.38) up with respect to k, we obtain (recall that q_x takes values in $\mathbb{R}^{d\times n}$)

$$\begin{aligned}
&-\mathbb{E}\int_{\mathbb{R}^n}|V_x^\varepsilon(t,x)|^2dx\\
&=\sum_{k=1}^n\mathbb{E}\int_t^T\int_{\mathbb{R}^n}\Big\{\left|\sigma^T\left(V_{x_k}^\varepsilon\right)_x\right|^2+2\left\langle q_{x_k}^\varepsilon,\sigma^T\left(V_{x_k}^\varepsilon\right)_x-V_{x_k}^\varepsilon\bar\sigma^0\right\rangle+|q_{x_k}^\varepsilon|^2-2V_{x_k}^\varepsilon\mathrm{tr}\left(\sigma_{x_k}q_x^\varepsilon\right)\\
&\quad+2\left\langle q^\varepsilon,\bar\sigma_{x_k}^0-\nabla\cdot\sigma_{x_k}\right\rangle V_{x_k}^\varepsilon+\left[\left(\nabla\cdot\bar b\right)-2\bar b^0\right]\left(V_{x_k}^\varepsilon\right)^2-2V_{x_k}^\varepsilon\bar g_{x_k}+2\psi'\left(\frac{V^\varepsilon}{\varepsilon}\right)\frac{\left(V_{x_k}^\varepsilon\right)^2}{\varepsilon}\\
&\quad-\frac{1}{2}\left\langle\left(\sigma\sigma^T\right)_{x_kx_k}V_x^\varepsilon,V_x^\varepsilon\right\rangle-2V_{x_k}^\varepsilon\left\langle\bar b_{x_k},V_x^\varepsilon\right\rangle+\bar b_{x_kx_k}^0\left(V^\varepsilon\right)^2\Big\}dxdr\\
&=\mathbb{E}\int_t^T\int_{\mathbb{R}^n}\Bigg\{\left|\sigma^TV_{xx}^\varepsilon\right|^2+2\,\mathrm{tr}\left[\left(q_x^\varepsilon\right)^T\sigma^TV_{xx}^\varepsilon\right]+|q_x^\varepsilon|^2\\
&\quad-2\,\mathrm{tr}\left[\left(q_x^\varepsilon\right)^T\left(\bar\sigma^0V_x^\varepsilon-\sum_{k=1}^nV_{x_k}^\varepsilon\sigma_{x_k}^T\right)\right]+2\left\langle q^\varepsilon,\left(\bar\sigma^0-\nabla\cdot\sigma\right)_xV_x^\varepsilon\right\rangle\\
&\quad+\ \left[\left(\nabla\cdot\bar b\right)-2\bar b^0\right]|V_x^\varepsilon|^2-2\left\langle V_x^\varepsilon,\bar g_x\right\rangle+2\psi'\left(\frac{V^\varepsilon}{\varepsilon}\right)\frac{|V_x^\varepsilon|^2}{\varepsilon}\\
&\quad-\frac{1}{2}\left\langle\left[\Delta\left(\sigma\sigma^T\right)\right]V_x^\varepsilon,V_x^\varepsilon\right\rangle-2\left\langle\bar b_xV_x^\varepsilon,V_x^\varepsilon\right\rangle+\left(\Delta\bar b^0\right)|V^\varepsilon|^2\Bigg\}dxdr\\
&=\mathbb{E}\int_t^T\int_{\mathbb{R}^n}\Bigg\{\left|\sigma^TV_{xx}^\varepsilon+q_x^\varepsilon-\bar\sigma^0V_x^\varepsilon-\sum_{k=1}^nV_{x_k}^\varepsilon\sigma_{x_k}^T\right|^2-\left|\bar\sigma^0V_x^\varepsilon+\sum_{k=1}^nV_{x_k}^\varepsilon\sigma_{x_k}^T\right|^2\\
&\quad+2\,\mathrm{tr}\left[V_{xx}^\varepsilon\left(\sigma\bar\sigma^0V_x^\varepsilon+\sum_{k=1}^nV_{x_k}^\varepsilon\sigma\sigma_{x_k}^T\right)\right]+2\left\langle q^\varepsilon,\left(\bar\sigma^0-\nabla\cdot\sigma\right)_xV_x^\varepsilon\right\rangle-2\left\langle V_x^\varepsilon,\bar g_x\right\rangle\\
&\quad+2\psi'\left(\frac{V^\varepsilon}{\varepsilon}\right)\frac{|V_x^\varepsilon|^2}{\varepsilon}-\Bigg\langle\left[\frac{1}{2}\Delta\left(\sigma\sigma^T\right)\right]+\bar b_x+\bar b_x^T-\left(\left(\nabla\cdot\bar b\right)-2\bar b^0\right)I]V_x^\varepsilon,V_x^\varepsilon\Bigg\rangle\\
&\quad+\left(\Delta\bar b^0\right)|V^\varepsilon|^2\Bigg\}dxdr.
\end{aligned}$$

Note that

$$\begin{aligned}
\mathbb{E}\int_t^T\int_{\mathbb{R}^n}\mathrm{tr}\left[V_{xx}^\varepsilon\sigma\bar\sigma^0\left(V_x^\varepsilon\right)^T\right]dxdr&=\mathbb{E}\int_t^T\int_{\mathbb{R}^n}\left(V_{xx}^\varepsilon V_x^\varepsilon\right)^T\sigma\bar\sigma^0dxdr\\
&=\frac{1}{2}\,\mathbb{E}\int_t^T\int_{\mathbb{R}^n}\left[\left(|V_x^\varepsilon|^2\right)_x\right]^T\sigma\bar\sigma^0dxdr\\
&=-\frac{1}{2}\,\mathbb{E}\int_t^T\int_{\mathbb{R}^n}|V_x^\varepsilon|^2\left(\nabla\cdot\left[\sigma\bar\sigma^0\right]\right)dxdr\\
&\geq-C\mathbb{E}\int_t^T\int_{\mathbb{R}^n}|V_x^\varepsilon|^2dxdr.
\end{aligned}\tag{7.39}$$

Also, we note that

$$\mathrm{tr}\left[V_{xx}^\varepsilon\sigma\sigma_{x_k}^T\right]=\mathrm{tr}\left[\left(V_{xx}^\varepsilon\sigma\sigma_{x_k}^T\right)^T\right]=\mathrm{tr}\left[\sigma_{x_k}\sigma^TV_{xx}^\varepsilon\right]=\mathrm{tr}\left[V_{xx}^\varepsilon\sigma_{x_k}\sigma^T\right].\tag{7.40}$$

Now, if we denote $\Phi_k = \sigma\sigma_{x_k}^T + \sigma_{x_k}\sigma^T$, then it is $\mathbb{R}^{n\times n}$-valued, symmetric, and

$$\begin{aligned}
\text{tr}\,\left[V_{xx}^\varepsilon V_{x_k}^\varepsilon \Phi_k\right] &= \text{tr}\,\left[\left(V_{x_k}^\varepsilon \Phi_k V_x^\varepsilon\right)_x\right] - \left\langle \nabla \cdot \left(V_{x_k}^\varepsilon \Phi_k\right), V_x^\varepsilon \right\rangle \\
&= \nabla \cdot \left(V_{x_k}^\varepsilon \Phi_k V_x^\varepsilon\right) - \langle\, V_{x_k}^\varepsilon \nabla \cdot \Phi_k, V_x^\varepsilon\,\rangle - \langle\, \Phi_k \left(V_x^\varepsilon\right)_{x_k}, V_x^\varepsilon\,\rangle \\
&= \nabla \cdot \left(V_{x_k}^\varepsilon \Phi_k V_x^\varepsilon\right) - \langle\, V_{x_k}^\varepsilon \nabla \cdot \Phi_k, V_x^\varepsilon\,\rangle - \frac{1}{2}\left[\langle\, \Phi_k V_x^\varepsilon, V_x^\varepsilon\,\rangle\right]_{x_k} \\
&\quad + \frac{1}{2}\,\langle(\Phi_k)_{x_k} V_x^\varepsilon, V_x^\varepsilon\,\rangle\,.
\end{aligned} \tag{7.41}$$

Thus,

$$\begin{aligned}
&\mathbb{E}\int_t^T \int_{\mathbb{R}^n} \text{tr}\,\left[V_{xx}^\varepsilon \sum_{k=1}^n V_{x_k}^\varepsilon \sigma\sigma_{x_k}^T\right] dxdr \\
&= \frac{1}{2}\mathbb{E}\int_t^T \int_{\mathbb{R}^n} \text{tr}\,\left[V_{xx}^\varepsilon \sum_{k=1}^n V_{x_k}^\varepsilon \left(\sigma\sigma_{x_k}^T + \sigma_{x_k}\sigma^T\right)\right] dxdr \\
&= \mathbb{E}\int_t^T \int_{\mathbb{R}^n} \left\{ -\frac{1}{2}\left\langle \sum_{k=1}^n V_{x_k}^\varepsilon \nabla \cdot \left[\sigma\sigma_{x_k}^T + \sigma_{x_k}\sigma^T\right], V_x^\varepsilon \right\rangle \right. \\
&\quad \left. + \frac{1}{4}\left\langle \left(\sum_{k=1}^n \left[\sigma\sigma_{x_k}^T + \sigma_{x_k}\sigma^T\right]_{x_k}\right) V_x^\varepsilon, V_x^\varepsilon \right\rangle \right\} dxdr \\
&\geq -\,C\mathbb{E}\int_t^T \int_{\mathbb{R}^n} |V_x^\varepsilon|^2 dxdr.
\end{aligned} \tag{7.42}$$

Consequently, making use of (5.18) with $m = 1$, we obtain

$$\begin{aligned}
-\mathbb{E}\int_{\mathbb{R}^n} |V_x^\varepsilon(t,x)|^2 dx \geq \mathbb{E}\int_t^T \int_{\mathbb{R}^n} &\left\{ \left| \sigma^T V_{xx}^\varepsilon + q_x^\varepsilon - \bar{\sigma}^0 (V_x^\varepsilon)^T - \sum_{k=1}^n V_{x_k}^\varepsilon \sigma_{x_k}^T \right|^2 - |q^\varepsilon|^2 \right. \\
&\left. + 2\psi'\left(\frac{V^\varepsilon}{\varepsilon}\right)\frac{|V_x^\varepsilon|^2}{\varepsilon} - C|V_x^\varepsilon|^2 - C|V^\varepsilon|^2 - |\bar{g}_x|^2 \right\} dxdr \\
\geq \mathbb{E}\int_t^T \int_{\mathbb{R}^n} &\left\{ \left|\sigma^T V_{xx}^\varepsilon + q_x^\varepsilon\right|^2 + 2\psi'\left(\frac{V^\varepsilon}{\varepsilon}\right)\frac{|V_x^\varepsilon|^2}{\varepsilon} \right. \\
&\left. - C|V_x^\varepsilon|^2 - |\bar{g}_x|^2 - C|\bar{g}|^2 \right\} dxdr.
\end{aligned}$$

By Gronwall's inequality, we obtain (5.22). □

Acknowledgment. The authors would like to thank the anonymous referee for his/her suggestive comments, which lead to the current much better version of the paper.

REFERENCES

[1] M. Beibel and H. R. Lerche, *Optimal stopping of regular diffusion under random discounting*, Theory Probab. Appl., 45 (2001), pp. 547–557.

[2] A. Bensoussan, *Stochastic Control by Functional Analysis Methods*, North-Holland, Amsterdam, 1982.

[3] J. M. Bismut and B. Skalli, *Temps d'arrêt, thérie générale de processus et processus de Markov*, Z. Wahrsch. verw. Gebiete, 39 (1977), pp. 301–313.

[4] R. Buckdahn and J. Ma, *Pathwise stochastic control problems and stochastic HJB equations*, SIAM J. Control Optim., 45 (2007), pp. 2224–2256.

[5] S. Dayanik and I. Karatzas, *On the optimal stopping problem for one-dimensional diffusions*, Stochastic Process. Appl., 107 (2003), pp. 173–212.

[6] C. Dellarcherie, *Capacités et Processus Stochastiques*, Springer-Verlag, New York, 1972.

[7] N. Egglezos and I. Karatzas, *Aspects of utility maximization with habit formation: Dynamic programming and stochastic PDE's*, preprint, Columbia University, New York, 2007.

[8] N. El Karoui, *Les aspects probabilistes du contrôle stochastique*, in Ecole d'Été de Probabilités de Saint-Flour IX-1979, Lecture Notes in Math. 876, Springer-Verlag, New York, pp. 73–238.

[9] A. G. Fakeev, *Optimal stopping rules for processes with continuous parameter*, Theory Probab. Appl., 15 (1970), pp. 324–331.

[10] A. G. Fakeev, *Optimal stopping of a Markov process*, Theory Probab. Appl., 16 (1971), pp. 694–696.

[11] R. Fernholz and I. Karatzas, *Stochastic portfolio theory: An overview*, preprint, Columbia University, New York, 2008.

[12] A. Friedman, *Variational Inequalities and Free Boundary Problems*, John Wiley & Sons, New York, 1983.

[13] I. Karatzas and S. Shreve, *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York, 1998.

[14] I. Karatzas and S. Shreve, *Methods of Mathematical Finance*, Springer-Verlag, New York, 1998.

[15] H. Kunita, *Stochastic Flows and Stochastic Differential Equations*, Cambridge University Press, London, 1990.

[16] J. Ma and J. Yong, *On linear, degenerate backward stochastic partial differential equations*, Probab. Theory Related Fields, 113 (1999), pp. 135–170.

[17] J. Ma and J. Yong, *Forward-Backward Stochastic Differential Equations and Their Applications*, Springer-Verlag, Berlin, 1999.

[18] J. Ma and J. Yong, *Dynamic programming for multidimensional stochastic control problems*, Acta Math. Sinica, 15 (1999), pp. 485–506.

[19] A. Nikeghbali, *An essay on the general theory of stochastic processes*, Prob. Surv., 3 (2006), pp. 345–412.

[20] E. Pardoux and A. Rascanu, *Backward stochastic variational inequalities*, Stoch. Stoch. Rep., 67 (1999), pp. 159–167.

[21] S. Peng, *Stochastic Hamilton–Jacobi–Bellman equations*, SIAM J. Control Optim., 30 (1992), pp. 284–304.

[22] S. Peng, *Backward stochastic differential equations and applications to optimal control*, Appl. Math. Optim., 27 (1993), pp. 125–144.

[23] G. Peskir and A. Shiryaev, *Optimal Stopping and Free-Boundary Problems*, Birkhäuser-Verlag, Basel, 2006.

[24] A. N. Shirayaev, *Optimal Stopping Rules*, Springer-Verlag, New York, 1978.

[25] M. E. Thompson, *Continuous parameter optimal stopping problems*, Z. Wahrsch. verw. Gebiete, 19 (1971), pp. 302–318.

[26] S. Tang, *Semi-linear systems of backward stochastic partial differential equations in* $\mathbb{R}^n$, Chinese Ann. Math. Ser. B, 26 (2005), pp. 437–456.

[27] S. Tang and S. H. Hou, *Switching games of stochastic differential systems*, SIAM J. Control Optim., 46 (2007), pp. 900–929.

[28] S. Tang and J. Yong, *Finite horizon stochastic optimal switching and impulse controls with a viscosity solution approach*, Stoch. Stoch. Rep., 45 (1993), pp. 145–176.

[29] J. Yong and X. Y. Zhou, *Stochastic Controls: Hamiltonian Systems and HJB Equations*, Springer-Verlag, New York, 1999.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 2, pp. 972–992

© 2009 Society for Industrial and Applied Mathematics

LOCAL EXACT CONTROLLABILITY AND STABILIZABILITY OF THE NONLINEAR SCHRÖDINGER EQUATION ON A BOUNDED INTERVAL[*]

LIONEL ROSIER[†] AND BING-YU ZHANG[‡]

Abstract. This paper studies the exact controllability and the stabilization of the cubic Schrödinger equation posed on a bounded interval. Both internal and boundary controls are considered, and the results are given first in a periodic setting, and next with Dirichlet (resp., Neumann) boundary conditions. It is shown that the systems with either an internal control or a boundary control are locally exactly controllable in the classical Sobolev space H^s for any $s \geq 0$. It is also shown that the systems with an internal stabilization are locally exponentially stabilizable in H^s for any $s \geq 0$.

Key words. Schrödinger equation, Bourgain spaces, exact boundary controllability, exact internal controllability, exponential stabilization

AMS subject classifications. 93B05, 93D15, 35Q53

DOI. 10.1137/070709578

1. Introduction.

In this paper we study the nonlinear Schrödinger equation

$$iu_t + \lambda|u|^2 u + u_{xx} = 0, \tag{1.1}$$

where $u = u(x,t)$ is a complex-valued function of two real variables x and t, the subscripts denote the corresponding partial derivatives, and the parameter λ is a nonzero real constant. The equation arises in various physical contexts as a model for propagation of nonlinear waves. In optics, it may serve as a model of wave propagation in fiber optics, the function u represents a wave, and the equation describes the propagation of the wave through a nonlinear medium. The equation is also used as a model for some water waves to describe the evolution of the envelope of modulated wave groups. The value of the nonlinearity parameter λ depends on the relative water depth. For deep water, with the water depth large compared to the wave length of the water waves, λ is positive and envelope solitons may occur [27].

Our main concern in this paper is control and stabilization of the system described by (1.1). Consideration will be first given to internal control of the Schrödinger equation

$$iu_t + \lambda|u|^2 u + u_{xx} = f \tag{1.2}$$

posed on the finite interval $(-\pi, \pi)$ with periodic boundary conditions

$$u(-\pi,t) = u(\pi,t), \qquad u_x(-\pi,t) = u_x(\pi,t), \tag{1.3}$$

or posed on the finite interval $(0,\pi)$ with either the Dirichlet boundary conditions

$$u(0,t) = 0, \qquad u(\pi,t) = 0, \tag{1.4}$$

[*] Received by the editors November 29, 2007; accepted for publication (in revised form) November 13, 2008; published electronically February 25, 2009.
http://www.siam.org/journals/sicon/48-2/70957.html

[†] Institut Elie Cartan, Université Henri Poincaré Nancy 1, B.P. 239, 54506 Vandœuvre-lès-Nancy Cedex, France (rosier@iecn.u-nancy.fr). Partially supported by INRIA-CONICYT.

[‡] Department of Mathematical Sciences, University of Cincinnati, Cincinnati, OH 45221 (bzhang@math.uc.edu).

or the Neumann boundary conditions

$$u_x(0,t) = 0, \qquad u_x(\pi,t) = 0. \tag{1.5}$$

Here $f = f(x,t)$ is a given function considered as a control input. Without loss of generality, we assume that λ takes the value of 1 or -1 and restrict our attention to controls of the form

$$f(x,t) = iGh := ig(x)h(x,t), \tag{1.6}$$

where g, called a controller, is a given nonzero real-valued smooth function with its support contained in the domain $(-\pi,\pi)$ in the case of periodic boundary conditions, or in the domain $(0,\pi)$ in the case of Dirichlet boundary conditions or Neumann boundary conditions, and $h(x,t)$ is a new control input.

Then we will turn to boundary control of the nonlinear Schrödinger equation

$$iu_t + u_{xx} + \lambda|u|^2u = 0 \tag{1.7}$$

posed on the finite interval $(0,\pi)$ with either the Dirichlet boundary conditions

$$u(0,t) = h(t), \qquad u(\pi,t) = 0 \tag{1.8}$$

or the Neumann boundary conditions

$$u_x(0,t) = h(t), \qquad u_x(\pi,t) = 0, \tag{1.9}$$

where the boundary value function h will be considered as a control input.

In this paper, the focus of our study is the following two control problems.

> **Exact controllability problem**: Let $T > 0$ be given. Given the initial state u_0 and the terminal state u_1 in an appropriate space, can one find a control h such that system (1.2)–(1.3) (resp., system (1.2)–(1.4) or system (1.2)–(1.5)) admits a solution $u(x,t)$ satisfying
>
> $$u(x,0) = u_0(x), \qquad u(x,T) = u_1(x)?$$
>
> **Stabilizability problem**: Can one find a linear feedback control law $h = Ku$ such that the resulting closed-loop system is exponentially stable?

Control and stabilization problems of the Schrödinger equation have received a lot of attention in the past decade.[1] While significant progresses have been made for the linear Schrödinger equation on its controllability and stabilizability properties (cf., e.g., [4, 10, 8, 11, 12, 13, 14, 15, 17, 16, 20, 21]), there are only a few results for the nonlinear Schrödinger equation. Illner, Lange, and Teismann [6, 7] considered internal controllability of the nonlinear Schrödinger equation posed on a finite interval with periodic boundary conditions. They showed that the system (1.2)–(1.3) is locally exactly controllable in the space $H^1_p(-\pi,\pi) := \{v \in H^1(-\pi,\pi) : \; v(-\pi) = v(\pi)\}$. Their approach is based on the well-known Hilbert uniqueness method (HUM) and Schauder's fixed point theorem. Later, Lange and Teismann [9] considered internal control of the nonlinear Schrödinger equation (1.2) posed on a finite interval with the homogeneous Dirichlet boundary conditions (1.4) and established local exact controllability of the system (1.2)–(1.4) in the space $H^1_0(0,\pi)$ around a special ground state of the system.

[1] The readers are referred to Zuazua [29] for an excellent review on recent progresses of this subject up to 2003.

Recently Dehman, Gérard, and Lebeau [5] studied internal control and stabilization of a class of defocusing nonlinear Schrödinger equations posed on a two-dimensional compact Riemannian manifold M without boundary. They demonstrated, in particular, that the system is semiglobally exact controllable and semiglobally exponentially stabilizable in the space $H^1(M)$ assuming both *the geometric control condition* and *the unique continuation property*[2] are satisfied.

There are two natural energy spaces associated with the nonlinear Schrödinger equation (1.2), namely $L^2(I)$ or $H^1(I)$. Here I stands for either the interval $(-\pi, \pi)$ or the interval $(0, \pi)$. Indeed, let u be a smooth solution of (1.2) with the control input $h \equiv 0$ satisfying one of the boundary conditions (1.3), (1.4), and (1.5). Then it satisfies the following two conservation laws:

$$E_0(t) := \int_I |u(x,t)|^2 dx = E_0(0)$$

and

$$E_1(t) := \int_I |u_x(x,t)|^2 dx - \frac{\lambda}{2} \int_I |u(x,t)|^4 dx = E_1(0)$$

for any $t \in \mathbb{R}$. While the local exact controllability of (1.2) has been established in the space $H^1(I)$ in [7] and [9], it would be interesting to know whether the nonlinear Schrödinger equation (1.2) is exactly controllable in the space $L^2(I)$ or in other Sobolev spaces $H^s(I)$ with $s \geq 0$.

One of our main objectives is to establish local exact controllability of (1.2) in the space $H^s(I)$ for any $s \geq 0$. In order to describe precisely our results, we introduce the following notations.

Let

$$\phi_k(x) = \frac{1}{\sqrt{2\pi}} e^{ikx} \qquad\qquad k = 0, \pm 1, \pm 2, \dots .$$

Then $\{\phi_k\}_{k=-\infty}^{+\infty}$ forms an orthonormal basis in the space $L^2(-\pi, \pi)$. We may define the Sobolev space $H_p^s := H_p^s(-\pi, \pi)$ of order s $(s \geq 0)$ as the space of all 2π-periodic functions

$$v(x) = \sum_{k=-\infty}^{\infty} v_k \phi_k(x)$$

such that

$$\left\{ \sum_{k=-\infty}^{\infty} |v_k|^2 (1 + |k|)^{2s} \right\}^{\frac{1}{2}} < \infty. \tag{1.10}$$

The left-hand side of (1.10) is a Hilbert norm for H_p^s; we denote it by $\|v\|_s$. In addition, let

$$C_{odd}^\infty(0, \pi) = \left\{ v \in C^\infty[0, \pi];\ v^{(2k)}(0) = v^{(2k)}(\pi) = 0, \quad k = 0, 1, 2, \dots \right\}$$

and

$$C_{even}^\infty(0, \pi) = \left\{ v \in C^\infty[0, \pi];\ v^{(2k+1)}(0) = v^{(2k+1)}(\pi) = 0, \quad k = 0, 1, 2, \dots \right\}.$$

[2] See [5] for exact descriptions of these two conditions.

Obviously, both $C^\infty_{odd}(0,\pi)$ and $C^\infty_{even}(0,\pi)$ are subspaces of $H^s(0,\pi)$ for any $s \geq 0$. Let $H^s_{odd}(0,\pi)$ and $H^s_{even}(0,\pi)$ be the closure of $C^\infty_{odd}(0,\pi)$ and $C^\infty_{even}(0,\pi)$ in the space $H^s(0,\pi)$, respectively. Note that $H^0_{odd}(0,\pi) = H^0_{even}(0,\pi) = L^2(0,\pi)$ and $H^1_{odd}(0,\pi) = H^1_0(0,\pi)$.

We have the following local controllability result for system (1.2)–(1.3).

Theorem 1.1. *Let $T > 0$ and $s \geq 0$ be given. There exists a $\delta > 0$ such that for any $u_0,\ u_1 \in H^s_p(-\pi,\pi)$ satisfying*

$$\|u_0\|_s \leq \delta, \qquad \|u_1\|_s \leq \delta,$$

there exists a control $h \in L^2([0,T]; H^s_p(-\pi,\pi))$ such that the system (1.2)–(1.3) *admits a solution $u \in C([0,T]; H^s_p(-\pi,\pi))$ satisfying*

$$u(x,0) = u_0(x), \qquad u(x,T) = u_1(x).$$

For the Schrödinger equation (1.2) posed on the finite interval $(0,\pi)$ with the Neumann boundary conditions (1.5), we have the following local controllability result.

Theorem 1.2. *Let $T > 0$ and $s \geq 0$ be given. There exists a $\delta > 0$ such that for any $u_0,\ u_1 \in H^s_{even}(0,\pi)$ satisfying*

$$\|u_0\|_{H^s(0,\pi)} \leq \delta, \qquad \|u_1\|_{H^s(0,\pi)} \leq \delta,$$

there exists a control $h \in L^2([0,T]; H^s(0,\pi))$ such that the system (1.2)–(1.5) *admits a solution $u \in C([0,T]; H^s_{even}(0,\pi))$ satisfying*

$$u(x,0) = u_0(x), \qquad u(x,T) = u_1(x).$$

A similar result holds for the Schrödinger equation (1.2) posed on the finite interval $(0,\pi)$ with the Dirichlet boundary conditions (1.4).

Theorem 1.3. *Let $T > 0$ and $s \geq 0$ be given. There exists a $\delta > 0$ such that for any $u_0,\ u_1 \in H^s_{odd}(0,\pi)$ satisfying*

$$\|u_0\|_{H^s(0,\pi)} \leq \delta, \qquad \|u_1\|_{H^s(0,\pi)} \leq \delta,$$

there exists a control $h \in L^2([0,T]; H^s(0,\pi))$ such that the system (1.2)–(1.4) *admits a solution $u \in C([0,T]; H^s_{odd}(0,\pi))$ satisfying*

$$u(x,0) = u_0(x), \qquad u(x,T) = u_1(x).$$

For $s \geq 0$, let $H^s_{e,\pi}$ be the closure of the set $\{v \in C^\infty[0,\pi]; v^{(2k)}(\pi) = 0, k = 0,1,2,\ldots\}$ in the space $H^s(0,\pi)$ and $H^s_{o,\pi}$ be the closure of the set $\{v \in C^\infty[0,\pi]; v^{(2k+1)}(\pi) = 0, k = 0,1,2,\cdots\}$ in the space $H^s(0,\pi)$ for any $s \geq 0$. Then we have the following boundary controllability results for the systems (1.7)–(1.8) and (1.7)–(1.9).

Theorem 1.4. *Let $s \geq 0$ and $T > 0$ be given. There exists a $\delta > 0$ such that*

(a) *for any $u_0,\ u_1 \in H^s_{e,\pi}(0,\pi)$ satisfying and*

$$\|u_0\|_{H^s(0,\pi)} \leq \delta, \qquad \|u_1\|_{H^s(0,\pi)} \leq \delta,$$

there exists a boundary control h such that the system (1.7)–(1.8) *admits a solution $u \in C([0,T]; H^s_{e,\pi}(0,\pi))$ satisfying*

$$u(x,0) = u_0(x), \qquad u(x,T) = u_1(x);$$

(b) *for any* $u_0,\ u_1 \in H^s_{o,\pi}(0,\pi)$ *satisfying*

$$\|u_0\|_{H^s(0,\pi)} \le \delta, \qquad \|u_1\|_{H^s(0,\pi)} \le \delta,$$

there exists a boundary control h *such that the system* (1.7)–(1.9) *admits a solution* $u \in C([0,T]; H^s_{o,\pi}(0,\pi))$ *satisfying*

$$u(x,0) = u_0(x), \qquad u(x,T) = u_1(x).$$

Remarks.

(i) The same results hold if we apply a boundary control h on the right end of the domain $x = \pi$.

(ii) As it will be demonstrated in the proof in section 3, the boundary control h is taken as the trace of a function $w \in C([0,T]; H^s(-\epsilon,\pi))$ at $x = 0$, which is a solution of the system

$$\begin{cases} iw_t + w_{xx} + \lambda|w|^2 w = ig(x)h(x,t), & x \in (-\epsilon,\pi), \\ w(-\epsilon,t) = w(\pi,t), \qquad w_x(-\epsilon,t) = w_x(\pi,t). \end{cases}$$

Our other main objective in this paper is to study stabilizability of system (1.2)–(1.3). We will show that it is locally stabilizable in $H^s_p(-\pi,\pi)$ $(s \ge 0)$ by the feedback law

$$h(x,t) = -g(x)u(x,t).$$

Theorem 1.5. *Let* $s \ge 0$ *be given. There exist some positive constants* δ, C, *and* ν *such that every solution of the system*

$$iu_t + u_{xx} + \lambda|u|^2 u = -ig^2(x)u, \qquad u(x,0) = u_0(x) \tag{1.11}$$

$$u(-\pi,t) = u(\pi,t), \qquad u_x(-\pi,t) = u(\pi,t) \tag{1.12}$$

issued from an initial state $u_0 \in H^s_p(-\pi,\pi)$ *with* $||u_0||_s \le \delta$ *satisfies*

$$||u(t)||_s \le Ce^{-\nu t}||u_0||_s \qquad \forall t \ge 0. \tag{1.13}$$

Remarks.

(i) In Theorem 1.5, the regularity of g may be weakened to $g \in L^\infty(-\pi,\pi)$ for $s = 0$, and to $g \in H^s_p(-\pi,\pi)$ for $s > 1/2$. The solution of (1.11)–(1.12) is proved to exist and to be unique in some restricted Bourgain space.

(ii) The L^2 norm of any solution of (1.11)–(1.12) is nondecreasing whatever be λ. Indeed, a simple computation gives

$$\int_{-\pi}^{\pi} |u(T,x)|^2 dx - \int_{-\pi}^{\pi} |u_0(x)|^2 dx = -2\int_0^T\int_{-\pi}^{\pi} |u(x,t)|^2 g(x)\, dx.$$

We conjecture that when $s = 0$ Theorem 1.5 is valid for any $\delta > 0$, i.e., that a semiglobal stabilization occurs in $L^2(-\pi,\pi)$.

(iii) Similar results may be derived for the systems with either Dirichlet or Neumann boundary conditions.

Our next theorem presents a local stabilization result for a general nonlinearity in the Sobolev space $H^s_p(-\pi,\pi)$ with $s > 1/2$.

THEOREM 1.6. *Let $s > 1/2$, and let $F : \mathbb{C} \to \mathbb{C}$ be a continuous function such that for some positive constants r, C, and N it holds*

$$||F(u) - F(v)||_s \leq C \left(||u||_s^N + ||v||_s^N \right) ||u - v||_s$$

for all $u, v \in H_p^s(-\pi, \pi)$ with $||u||_s < r$, $||v||_s < r$. For any $\mu > 0$, let B_μ denote the space

$$B_\mu = \left\{ u \in C\left(\mathbb{R}^+; H_p^s(-\pi,\pi)\right);\ \left\| e^{\mu t} u(t) \right\|_{L^\infty(\mathbb{R}^+, H_p^s(-\pi,\pi))} < \infty \right\}$$

endowed with its natural norm. Let the function g be as in Theorem 1.5. *Then there exist positive constants δ, μ, and K such that for any $u_0 \in H_p^s(-\pi, \pi)$ with $||u_0||_s < \delta$, the system*

$$iu_t + u_{xx} + F(u) = -ig^2(x)u, \qquad u(x,0) = u_0(x) \tag{1.14}$$

$$u(-\pi, t) = u(\pi, t), \qquad u_x(-\pi, t) = u(\pi, t) \tag{1.15}$$

admits a unique solution $u \in B_\mu$, and it holds

$$||u(t)||_s \leq K e^{-\mu t} ||u_0||_s \qquad \forall t \geq 0.$$

To prove the above theorems we use the approach developed earlier by Russell and Zhang [25] in dealing with control and stabilization problem of the Korteweg–de Vries equation posed on a periodic domain. The associated linear systems are studied first using the classical moment method which enables us to establish the exact controllability of the associated linear systems in the space $H^s(I)$ for any $s \geq 0$. The linear results are then extended to the nonlinear systems. During this process, the Bourgain smoothing property [2, 3] for solutions of the Schrödinger equation posed on a periodic domain plays a key role. In particular, this Bourgain smoothing property seems indispensable in establishing the exact controllability of the system (1.2)–(1.3) in the space $L^2(-\pi, \pi)$. The proof of exact controllability for systems (1.2)–(1.4) and (1.2)–(1.5) is based on the following observation (see [1] for more detail and its application in establishing well-posedness of nonhomogeneous boundary value problems of the nonlinear Schrödinger equation posed on a bounded domain):

If $u \in C([0, T]; H_p^s(-\pi, \pi))$ is an odd (even) function with respect to x-variable and solves (1.1), *then its restriction $w = w(x, t)$ on the interval* $(0, \pi)$ *belongs to the space $C([0, T]; H^s(0, \pi))$ and is a solution of the system* (1.1)–(1.4) *(system* (1.1)–(1.5)). *On the other hand, if $w \in C([0, T]; H^s(0, \pi))$ solves system* (1.1)–(1.4) *(or system* (1.1)–(1.5)*) and u is its odd (or even) extension to the interval $(-\pi, \pi)$, then $u \in C([0, T]; H_p^s(-\pi, \pi))$ and solves system* (1.1)–(1.3).

Thus, one can reduce exact control problem of systems (1.2)–(1.4) and (1.2)–(1.5) to that of system (1.2)–(1.3). Theorem 1.2 and Theorem 1.3 can be considered as corollaries of Theorem 1.1. As for boundary control systems (1.7)–(1.8) and (1.7)–(1.9), their exact controllability follows from internal controllability of systems (1.2)–(1.4) and (1.2)–(1.5) by a standard procedure.

The paper is outlined as follows. In section 2, we establish the exact internal controllability of the linear Schrödinger equation with periodic boundary conditions by using the moment approach. In section 3, we derive the (internal or boundary) exact controllability of the cubic Schrödinger equation with various boundary conditions. The internal stabilization is investigated in section 4. In particular, the proof of

Theorem 1.5 is presented in this section. As for Theorem 1.6, its proof is similar to the one in [18, Theorem 1.1] and is, therefore, omitted.

Finally, the readers are referred to [19, 22, 23, 24, 25, 28] and references therein for the study of control and stabilization of another important nonlinear dispersive wave equation, the Korteweg–de Vries equation.

2. Linear systems. We first consider the associated linear open loop control system of the Schrödinger equation posed on $(-\pi,\pi)$ with the periodic boundary conditions:

$$\begin{cases} iv_t + v_{xx} = iGh, & v(x,0) = v_0(x), \\ v(-\pi,t) = v(\pi,t), & v_x(-\pi,t) = v_x(\pi,t), \end{cases} \tag{2.1}$$

where the operator G is defined by (1.6) and h is the applied control function.

Let A denote the operator

$$Aw = iw'' \tag{2.2}$$

on the domain $\mathcal{D}(A) = H^2_p(-\pi,\pi)$. It generates a strongly continuous group $W(t)$ in the space $L^2(-\pi,\pi)$; the eigenfunctions are simply the orthonormal Fourier basis functions in $L^2(-\pi,\pi)$

$$\phi_k(x) = \frac{1}{\sqrt{2\pi}}e^{ikx}, \qquad k = 0, \pm1, \pm2, \ldots .$$

We have the following exact controllability result for the system (2.1).

Theorem 2.1. *Let $T > 0$ and $s \geq 0$ be given. For any v_0, $v_1 \in H^s_p(-\pi,\pi)$, there exists a control $h \in L^2([0,T]; H^s_p(-\pi,\pi))$ such that the system* (2.1) *admits a unique solution*

$$v \in C\left([0,T]; H^s_p(-\pi,\pi)\right)$$

satisfying

$$v(x,T) \equiv v_1(x).$$

Proof. The system (2.1) can be rewritten as an abstract control system in the space $H^s_p(-\pi,\pi)$,

$$\frac{d}{dt}v(t) = Av(t) + Gh, \qquad v(0) = v_0. \tag{2.3}$$

By the standard semigroup theory, for any $s \geq 0$, $T > 0$, $v_0 \in H^s_p(-\pi,\pi)$, and $h \in L^2(0,T; H^s_p(-\pi,\pi))$, (2.2) admits a unique solution $v \in C([0,T]; H^s_p(-\pi,\pi))$. It is familiar that the operator A, as defined in (2.2), has eigenvalues

$$\lambda_k = -ik^2$$

with the corresponding eigenfunctions ϕ_k for $-\infty < k < \infty$. Relative to the basis $\{\phi_k\}_{-\infty}^{\infty}$, the initial state v_0 and the terminal state v_1 have the expansions, convergent in $H^s_p(-\pi,\pi)$,

$$v_j = \sum_{k=-\infty}^{\infty} v_{k,j}\phi_k, \qquad v_{k,j} = \int_{-\pi}^{\pi} v_j(x)\overline{\phi_k(x)}dx \quad for\ j = 0, 1 \tag{2.4}$$

and the solution v has the expansion

$$v(x,t) = \sum_{k=-\infty}^{+\infty} v_{k,0} e^{\lambda_k t} \phi_k(x) + \sum_{k=-\infty}^{+\infty} \int_0^t e^{\lambda_k (t-\tau)} (Gh)_k(\tau) d\tau \phi_k(x),$$

where

$$(Gh)_k(t) = \int_{-\pi}^{\pi} g(x) h(x,t) \overline{\phi_k(x)} dx, \quad k = 0, \pm 1, \pm 2, \dots .$$

In order to find an appropriate control input h such that $v(x,T) = v_1(x)$, it suffices to solve the following moment problem:

$$v_{k,1} - v_{k,0} e^{\lambda_k T} = \int_0^T e^{\lambda_k (T-\tau)} (Gh)_k(\tau) d\tau \tag{2.5}$$

for $k = 0, \pm 1, \dots$.

Defining $p_k(t) = e^{\lambda_k t}$, $\mathcal{P} \equiv \{p_k \mid 0 \le k < \infty\}$ may be seen, from the result in [26], to form a Riesz basis for its closed span, $\mathcal{P}_T$, in $L^2(0,T)$. We let $\mathcal{Q} \equiv \{q_k \mid 0 \le k < \infty\}$ be the unique dual Riesz basis for $\mathcal{P}$ in $\mathcal{P}_T$, which fulfills

$$\int_0^T q_j(t) \overline{p_k(t)} dt = \delta_{kj}, \qquad 0 \le j,\, k < \infty. \tag{2.6}$$

We take the control h in (2.1) to have the form

$$h(x,t) = \sum_{j=-\infty}^{+\infty} h_j q_j(t) (G\phi_j)(x), \tag{2.7}$$

where $q_{-j} = q_j$ for $j \ge 0$ and the coefficients h_j are to be determined so that, among other things, the series (2.7) is appropriately convergent. Substituting (2.7) into (2.5) yields, using the biorthogonality (2.6),

$$v_{0,1} - v_{0,0} = \sum_{j=-\infty}^{+\infty} h_j \int_0^T e^{\overline{\lambda_0 t}} q_j(t) \int_{-\pi}^{\pi} G(G\phi_j)(x) \overline{\phi_0(x)} dx dt = h_0 \int_{-\pi}^{\pi} |(G\phi_0)(x)|^2 dx \tag{2.8}$$

and for $-\infty < k < \infty$, $k \neq 0$,

$$\begin{aligned} v_{k,1} - v_{k,0} e^{\lambda_k T} &= e^{\lambda_k T} \sum_{j=-\infty}^{+\infty} h_j \int_0^T e^{\overline{\lambda_k t}} q_j(t) \int_{-\pi}^{\pi} G(G\phi_j)(x) \overline{\phi_k(x)} dx dt \\ &= h_k e^{\lambda_k T} \int_{-\pi}^{\pi} |(G\phi_k)(x)|^2 dx + h_{-k} e^{\lambda_k T} \int_{-\pi}^{\pi} G\phi_{-k}(x) \overline{G\phi_k(x)} dx \end{aligned} \tag{2.9}$$

as G is a self-adjoint operator in $L^2(-\pi,\pi)$. Since

$$\int_{-\pi}^{\pi} |G\phi_k(x)|^2 dx = \frac{1}{2\pi} \int_{-\pi}^{\pi} g^2(x) dx := a,$$

$$\int_{-\pi}^{\pi} G\phi_{-k}(x) \overline{G\phi_k(x)} dx = \frac{1}{2\pi} \int_{-\pi}^{\pi} g^2(x) e^{-2ikx} dx := b_k$$

for $-\infty < k < \infty$, (2.8)–(2.9) may be rewritten as

$$c_0 = h_0 a \tag{2.10}$$

$$\begin{cases} c_k = ah_k + b_k h_{-k} \\ c_{-k} = b_{-k} h_k + ah_{-k} \end{cases} \qquad k = 1, 2, \ldots, \tag{2.11}$$

where $c_k = v_{k,1} e^{-\lambda_k T} - v_{k,0}$. As $\beta_k := a^2 - b_k b_{-k} \neq 0$ for any $k > 0$ and

$$\lim_{k\to\infty} |\beta_k| = a^2,$$

there exists a $\gamma > 0$ such that

$$|\beta_k| > \gamma \qquad \forall\ k > 0.$$

Thus, it follows from (2.10)–(2.11) by the Gram's rule that

$$h_0 = a^{-1} c_0, \qquad h_k = \beta_k^{-1}(ac_k - c_{-k} b_k), \quad k = \pm 1, \pm 2, \ldots. \tag{2.12}$$

It remains to show that h defined by (2.7) and (2.12) belongs to $L^2([0,T]; H_p^s(-\pi,\pi))$ provided that $v_0,\ v_1 \in H_p^s(-\pi,\pi)$. To this end, let us write

$$G\phi_j(x) = \sum_{k=-\infty}^{+\infty} a_{jk} \phi_k(x), \tag{2.13}$$

where

$$a_{jk} = \int_{-\pi}^{\pi} G\phi_j(x) \overline{\phi_k(x)} dx, \qquad -\infty < j,\ k < \infty.$$

Thus,

$$h(x,t) = \sum_{j=-\infty}^{+\infty} \sum_{k=-\infty}^{+\infty} h_j a_{jk} q_j(t) \phi_k(x)$$

and

$$\begin{aligned} \|h\|^2_{L^2([0,T];H_p^s(-\pi,\pi))} &= \int_0^T \sum_{k=-\infty}^{+\infty} (1+|k|)^{2s} \left| \sum_{j=-\infty}^{+\infty} a_{jk} h_j q_j(t) \right|^2 dt \\ &= \sum_{k=-\infty}^{+\infty} (1+|k|)^{2s} \int_0^T \left| \sum_{j=-\infty}^{+\infty} a_{jk} h_j q_j(t) \right|^2 dt \\ &\leq c \sum_{k=-\infty}^{+\infty} (1+|k|)^{2s} \sum_{j=-\infty}^{+\infty} |h_j|^2 |a_{jk}|^2 \\ &\leq c \sum_{j=-\infty}^{+\infty} |h_j|^2 \sum_{k=-\infty}^{+\infty} (1+|k|)^{2s} |a_{jk}|^2, \end{aligned} \tag{2.14}$$

where the constant c comes from the Riesz basis property of $\mathcal{Q}$ in $\mathcal{P}_T$. However,

$$|a_{jk}| = \left|(G\phi_j, \phi_k)_{L^2(-\pi,\pi)}\right| = \left|(g\phi_j, \phi_k)_{L^2(-\pi,\pi)}\right| = \left| \frac{1}{\sqrt{2\pi}} g_{k-j} \right|,$$

where

$$g = \sum_{m=-\infty}^{+\infty} g_m \phi_m.$$

Hence,

$$|a_{jk}|^2 \leq c|g_{k-j}|^2$$

and

$$\begin{aligned}\sum_{k=-\infty}^{+\infty} (1+|k|)^{2s}|a_{jk}|^2 &\leq c \sum_{k=-\infty}^{+\infty} (1+|k|)^{2s}|g_{k-j}|^2 \\ &\leq c \sum_{k=-\infty}^{+\infty} (1+|k+j|)^{2s}|g_k|^2 \\ &\leq c(1+|j|)^{2s} \sum_{k=-\infty}^{+\infty} (1+|k|)^{2s}|g_k|^2 \\ &\leq c(1+|j|)^{2s}\|g\|_s^2.\end{aligned}$$

We have by (2.14)

$$\begin{aligned}\|h\|^2_{L^2([0,T];H^s_p(-\pi,\pi))} &\leq c\left(\sum_{j=-\infty}^{+\infty}(1+|j|)^{2s}|h_j|^2\right)\|g\|_s^2 \\ &\leq c\left(\sum_{j=-\infty}^{+\infty}(1+|j|)^{2s}\frac{\left|e^{-\lambda_j T}v_{j,1}-v_{j,0}\right|^2}{|\beta_j|^2}\right)\|g\|_s^2 \\ &\leq c\sup_{j\neq 0}|\beta_j|^{-2}\|g\|_s^2\sum_{j=-\infty}^{+\infty}(1+|j|)^{2s}\left(|v_{j,1}|^2+|v_{j,0}|^2\right) \\ &\leq c\sup_{j\neq 0}\frac{1}{|\beta_j|^2}\|g\|_s^2\left(\|v_1\|_s^2+\|v_0\|_s^2\right).\end{aligned}$$

With this the proof is complete. □

Corollary 2.1. *Equations* (2.5), (2.7), *and* (2.12) *define, for $s \geq 0$, a bounded operator Φ :*

$$h = \Phi(v_0, v_1), \qquad \forall v_0,\ v_1 \in H^s_p(-\pi,\pi)$$

from $H^s_p(-\pi,\pi)\times H^s_p(-\pi,\pi)$ to $L^2([0,T];H^s_p(-\pi,\pi))$ such that

$$W(T)v_0 + \int_0^T W(T-\tau)(G(\Phi(v_0,v_1)))(\cdot,\tau)d\tau = v_1 \tag{2.15}$$

for any $(v_0,v_1)\in H^s_p(-\pi,\pi)\times H^s_p(-\pi,\pi)$ and

$$\|\Phi(v_0,v_1)\|_{L^2([0,T];H^s_p(-\pi,\pi))} \leq c(\|v_0\|_s + \|v_1\|_s), \tag{2.16}$$

where c depends only on T and g.

The following observation, while simple, is important to study the control properties of the linear Schrödinger equation posed on the finite interval with either Dirichlet boundary conditions or Neumann boundary conditions.

COROLLARY 2.2. *Assume that the control structure function g is an even function. If both the initial state v_0 and the terminal state v_1 are even (odd) functions of the x variable, then the control input h constructed in the proof of Theorem* 2.1 *is also an even (odd) function of the x variable and so is the corresponding solution v.*

Now we consider the linear Schrödinger equation posed on the finite interval $(0,\pi)$

$$iu_t + u_{xx} = iQh(x,t), \qquad u(x,0) = u_0(x), \qquad x \in (0,\pi),\ t \geq 0 \tag{2.17}$$

with either the Dirichlet boundary conditions

$$u(0,t) = 0, \qquad u(\pi,t) = 0 \tag{2.18}$$

or the Neumann boundary conditions

$$u_x(0,t) = 0, \qquad u_x(\pi,t) = 0, \tag{2.19}$$

where again $Qh(x,t) = q(x)h(x,t)$, $q(x)$ is a given smooth function supported in a subinterval of $(0,\pi)$.

THEOREM 2.2. *Let $s \geq 0$ and $T > 0$ be given. For any $u_0,\ u_1 \in H^s_{odd}(0,\pi)$, there exists*

$$h \in L^2(0,T;H^s(0,\pi))$$

such that (2.17)–(2.18) *admits a unique solution $u \in C([0,T];H^s(0,\pi))$ satisfying*

$$u(x,0) = u_0(x), \qquad u(x,T) = u_1(x), \qquad x \in (0,\pi).$$

THEOREM 2.3. *Let $s \geq 0$ and $T > 0$ be given. For any $u_0,\ u_1 \in H^s_{even}(0,\pi)$, there exists*

$$h \in L^2(0,T;H^s(0,\pi))$$

such that (2.17)–(2.19) *admits a unique solution $u \in C([0,T];H^s(0,\pi))$ satisfying*

$$u(x,0) = u_0(x), \qquad u(x,T) = u_1(x), \qquad x \in (0,\pi).$$

We provide just the proof of Theorem 2.2. The proof of Theorem 2.3 is similar and is, therefore, omitted.

Proof of Theorem 2.2. Note that if $u_0,\ u_1 \in H^s_{odd}(0,\pi)$, let v_0 and v_1 be the odd extension of u_0 and u_1, respectively, from $(0,\pi)$ to $(-\pi,\pi)$, then both v_0 and v_1 belong to the space $H^s_p(-\pi,\pi)$. Let g be the even extension of q from $(0,\pi)$ to $(-\pi,\pi)$ and consider the control system (2.1)–(2.2). According to Corollary 2.2, the corresponding control input $h(x,t)$ is also an odd (or even) function in the x variable. Consequently, the corresponding solution $v(x,t)$ is an odd function in the x variable, which, when restricted to the interval $(0,\pi)$, is a solution of the IBVP (2.17)–(2.18). The controllability results regarding (2.17)–(2.18) as described in Theorem 2.3, thus, follow from Theorem 2.1. The proof is complete. □

3. Exact controllability for NLS. In this section, we intend to extend the controllability results obtained for the linear Schrödinger equation to the nonlinear Schrödinger equation.

Consideration is first given to the system described by the nonlinear Schrödinger equation posed on the interval $(-\pi,\pi)$ with the periodic boundary conditions:

$$\begin{cases} iu_t + u_{xx} + \lambda|u|^2u = iGh, \qquad x\in(-\pi,\pi),\ t\in(0,T), \\ u(x,0)=u_0(x), \qquad u(-\pi,t)=u(\pi,t), \qquad u_x(-\pi,t)=u_x(\pi,t). \end{cases} \tag{3.1}$$

According to Bourgain [2, 3], for given $s\geq 0$, $u_0\in H_p^s(-\pi,\pi)$, and $h\in L^1_{loc}(\mathbb{R};H_p^s(-\pi,\pi))$, (3.1) admits a unique solution $u\in C(\mathbb{R};H_p^s(-\pi,\pi))$. Our main concern is the exact controllability of (3.1) as a distributive control system.

Recall that $W(t)$ is the C^0-group generated by the operator A, defined by (2.2), on the space $L^2(-\pi,\pi)$, with which, the system (3.1) has the following equivalent integral equation form:

$$u(t) = W(t)u_0 + \int_0^t W(t-\tau)(Gh)(\tau)d\tau + i\lambda\int_0^t W(t-\tau)(|u|^2u)(\tau)d\tau. \tag{3.2}$$

As it has been pointed out in the introduction, a smoothing property is needed for the operator from f to v:

$$v(t) = \int_0^t W(t-\tau)f(\tau)d\tau.$$

This needed smoothing property was provided in Bourgain's work [2, 3] where he dealt with the Cauchy problem for the periodic Schrödinger equation.

Let b and s be two given real numbers. For a function $w:\ (-\pi,\pi)\times R$, define the quantity

$$\Lambda_{b,s}(w) = \left(\sum_{n=-\infty}^{\infty}(1+|n|)^{2s}\int_{-\infty}^{\infty}|\hat{w}(n,\lambda)|^2(1+|\lambda+n^2|)^{2b}d\lambda\right)^{1/2},$$

where $\widehat{w}(n,\lambda)$ denotes the Fourier transform of $w(x,t)$ with respect to both the x and t variables. Following Bourgain [2, 3], we introduce the following space:

$$X^{b,s} = \left\{w\in L^2(\mathbb{R};H_p^s(-\pi,\pi));\ \Lambda_{b,s}(w)<\infty\right\} \tag{3.3}$$

with the norm $\|\cdot\|_{X^{b,s}} := \Lambda_{b,s}(\cdot)$. For any given $T>0$, $X_T^{b,s}$ denotes the restriction space of $X^{b,s}$ on the time interval $(0,T)$ with the associated quotient norm. It is clear that $X_T^{b,s}$ is a Hilbert space,

$$X_T^{0,s} = L^2(0,T;H_p^s(-\pi,\pi))$$

and

$$X_T^{b_1,s}\subset X_T^{b_2,s}$$

if $b_1>b_2$.

Before we proceed to show the exact controllability results, we present the following technical lemmas due to Bourgain [2, 3] which will play important roles in the proof of Theorem 1.1.

Lemma 3.1. *Let $T>0$, $s\geq 0$, and $0\leq b\leq 1$ be given. There exists a constant $C>0$ depending only on s and b such that*

$$\|W(t)\phi\|_{X_T^{b,s}}\leq C\|\phi\|_s$$

for any $\phi\in H_p^s(-\pi,\pi)$.

Lemma 3.2. *Let $T>0$, $s\geq 0$, and $b\in(\frac{3}{8},\frac{5}{8})$ and $\frac{5}{8}>b'>\max\{\frac{1}{2},b\}$ be given. There exists a constant $C>0$ depending only on s, b, and b' such that*

$$\left\|\int_0^t W(t-\tau)f(\tau)d\tau\right\|_{X_T^{b,s}}\leq C\|f\|_{X_T^{b'-1,s}}$$

for any $f\in X_T^{b'-1,s}$.

Lemma 3.3. *Let $s\geq 0$ and $b\in(\frac{3}{8},\frac{5}{8})$ and $\frac{5}{8}>b'>\max\{\frac{1}{2},b\}$ be given. There exist some constants $C>0$ and $\alpha>0$ depending only on s, b, and b' such that*

$$\|u\overline{v}w\|_{X_T^{b'-1,s}}\leq CT^{\alpha}\|u\|_{X_T^{b,s}}\|v\|_{X_T^{b,s}}\|w\|_{X_T^{b,s}}$$

for any $T>0$ and $u,\ v,\ w\in X_T^{b,s}$.

Now we are in a position to prove Theorem 1.1.

Proof of Theorem 1.1. Define

$$\omega(T,u):=-i\lambda\int_0^T W(T-\tau)\left(|u|^2u\right)(\tau)d\tau. \tag{3.4}$$

According to Corollary 2.1, for given $u_0,\ u_1\in H_p^s(-\pi,\pi)$, if one chooses

$$h=\Phi(u_0,u_1+\omega(T,u)) \tag{3.5}$$

in (3.2), then

$$\begin{aligned}u(t)=W(t)u_0&+\int_0^t W(t-\tau)(G\Phi(u_0,u_1+\omega(T,u)))(\tau)d\tau\\&+i\lambda\int_0^t W(t-\tau)\left(|u|^2u\right)(\tau)d\tau\end{aligned} \tag{3.6}$$

and

$$u(0)=u_0,\qquad u(T)=u_1 \tag{3.7}$$

by virtue of the definition of the operator Φ (cf. Corollary 2.1). This suggests that we consider the map

$$\begin{aligned}\Gamma(u)=W(t)u_0&+\int_0^t W(t-\tau)(G(\Phi(u_0,u_1+\omega(T,u))))(\tau)d\tau\\&+i\lambda\int_0^t W(t-\tau)\left(|u|^2u\right)(\tau)d\tau.\end{aligned} \tag{3.8}$$

If the map Γ is shown to be a contraction in an appropriate space, then its fixed point u is a solution of (1.2)–(1.3) with $h=\Phi(u_0,u_1+\omega(T,u))$ and satisfies $u(x,T)\equiv u_1(x)$. We show this is the case in the space $X_T^{b,s}$.

Note that $X_T^{0,s} = L^2(0,T;H_p^s(-\pi,\pi))$ and $X_T^{b_1,s}$ is continuously imbedded into $X_T^{b_2,s}$ if $b_2 < b_1$. Applying Lemma 3.1, Lemma 3.2, and Lemma 3.3 with $\frac{5}{8} > b' > b > \frac{1}{2}$ to (3.8) yields

$$\|\Gamma(u)\|_{X_T^{b,s}} \le c\|u_0\|_s + c\|G(\Phi(u_0, u_1 + \omega(T,u)))\|_{X_T^{b'-1,s}} + c\left\||u|^2u\right\|_{X_T^{b'-1,s}}.$$

It follows from the definitions of the operator Φ that

$$\begin{aligned}\|G(\Phi(u_0, u_1 + \omega(T,u)))\|_{X_T^{b'-1,s}} &\le c\,\|G\Phi(u_0, u_1 + \omega(T,u))\|_{L^2(0,T;H_p^s)}\\ &\le c\left(\|u_0\|_s + \|u_1\|_s + \|\omega(T,u)\|_s\right).\end{aligned}$$

Using Lemmas 3.1, 3.2, and 3.3, it follows that (note that $\lambda = 1$ or -1)

$$\begin{aligned}\|\omega(T,u)\|_s &= \|\int_0^T W(T-\tau)(|u|^2u)(\tau)d\tau\|_s\\ &\le \sup_{t\in[0,T]} \|\int_0^t W(t-\tau)(|u|^2u)(\tau)d\tau\|_s\\ &\le \|\int_0^t W(t-\tau)(|u|^2u)(\tau)d\tau\|_{X_T^{b,s}}\\ &\le c\||u|^2u\|_{X_T^{b'-1,s}}\\ &\le c\|u\|^3_{X_T^{b,s}}\end{aligned}$$

Consequently,

$$\|\Gamma(u)\|_{X_T^{b,s}} \le c\left(\|u_0\|_s + \|u_1\|_s\right) + c\,\|u\|^3_{X_T^{b,s}}. \tag{3.9}$$

For $M > 0$, let S_M be a bounded subset of X_s:

$$S_M = \left\{v \in X_T^{b,s},\ \|v\|_{X_T^{b,s}} \le M\right\}.$$

Then, for any $u \in S_M$,

$$\|\Gamma(u)\|_{X_T^{b,s}} \le c\|u_0\|_s + c\|u_1\|_s + cM^3.$$

We choose $\delta > 0$ and $M > 0$ such that

$$2c\delta + cM^3 \le M, \qquad cM^2 < 1/2. \tag{3.10}$$

Then,

$$\|\Gamma(u)\|_{X_T^{b,s}} \le M,$$

for any $u \in S_M$, if $\|u_0\|_s \le \delta$ and $\|u_1\|_s \le \delta$. In addition, for any $u,\ v \in S_M$, since

$$\begin{aligned}\Gamma(u) - \Gamma(v) = &\int_0^t W(t-\tau)(G\Phi(0, \omega(T,u) - \omega(T,v)))(\tau)d\tau\\ &+ i\lambda\int_0^t W(t-\tau)\left(|u|^2(u-v) + v\overline{u}(u-v) + (\overline{u}-\overline{v})v^2\right)(\tau)d\tau\end{aligned}$$

and

$$\omega(T,u)-\omega(T,v)=\int_0^T W(T-\tau)\left(|u|^2(u-v)+v\overline{u}(u-v)+(\overline{u}-\overline{v})v^2\right)(\tau)d\tau,$$

a similar argument shows that

$$\begin{aligned}\|\Gamma(u)-\Gamma(v)\|_{X_T^{b,s}} &\le c(\|u\|_{X_T^{b,s}}+\|v\|_{X_T^{b,s}})^2\|u-v\|_{X_T^{b,s}}\\ &\le cM^2\|u-v\|_{X_T^{b,s}}\\ &\le \frac{1}{2}\|u-v\|_{X_T^{b,s}}.\end{aligned}$$

Thus, the map Γ is a contraction on S_M provided that δ and M are chosen according to (3.10) and $\|u_0\|_s \le \delta$, $\|u_1\|_s \le \delta$. As a result, its fixed point $u \in S_M$ is the unique solution of the integral equation (3.8). The proof is complete. □

Next our attention is turned to the system described by the nonlinear Schrödinger equation posed on the finite interval $(0,\pi)$ with the Dirichlet boundary conditions:

$$\left\{\begin{array}{l} iv_t+v_{xx}+\lambda|v|^2v=Qh_1, \quad x\in(0,\pi),\ t\ge 0,\\ v(x,0)=v_0(x), \qquad v(0,t)=0, \qquad v(\pi,t)=0.\end{array}\right. \tag{3.11}$$

It has been shown in [1] that for given $s \ge 0$ and $v_0 \in H^s_{odd}(0,\pi)$ and $h_1 \in L^1_{loc}(\mathbb{R};H^s(0,\pi))$, (3.11) admits a unique solution $v \in C(\mathbb{R};H^s(0,\pi))$. Moreover, if we let g be the even extension of q from the interval $(0,\pi)$ to the interval $(-\pi,\pi)$, and u_0 be the odd extension of v_0 from the interval $(0,\pi)$ to the interval $(-\pi,\pi)$. If $u=u(x,t)$ is the odd extension solution $v(x,t)$ of (3.11) from the interval $(0,\pi)$ to the interval $(-\pi,\pi)$ with respect to the x-variable, then $u \in C(\mathbb{R};H^s_p(-\pi,\pi))$ and solves (3.1) with h being the odd extension of h_1 from the interval $(0,\pi)$ to the interval $(-\pi,\pi)$ with respect to the x-variable. On the other hand, if g is an even function, $u_0 \in H^s_p(-\pi,\pi)$ is an odd function and h is also an odd function with respect to the x-variable in (3.1), then the corresponding solution u of (3.1) is also an odd function. If we let $v(x,t)$ be its restriction on the interval $(0,\pi)$, then v solves (3.11) with v_0, q and h_1 being the restrictions of u_0, g, and h on the interval $(0,\pi)$, respectively.

Proof of Theorem 1.2. For given v_0 and $v_1 \in H^s_{odd}(0,\pi)$, let u_0 and u_1 be their odd extensions from $(0,\pi)$ to $(-\pi,\pi)$. We have $u_0,\ u_1 \in H^s_p(-\pi,\pi)$. In addition, let g be the even extension of q from $(0,\pi)$ to $(-\pi,\pi)$. It is sufficient to show that there exists a control input $h \in L^2(0,T;H^s_p(-\pi,\pi))$, which is odd with respect to the x-variable, such that (3.1) admits a solution $u \in C([0,T];H^s_p(-\pi,\pi))$ which is odd with respect to the x-variable and satisfies

$$u(x,T)=u_1(x).$$

Indeed, if this is the case, let v be the restriction of u to $(0,\pi)$ with respect to the x-variable. Then $v \in C([0,T];H^s(0,\pi))$ solves (3.11) and satisfies

$$v(x,T)=v_1(x).$$

To this end, as in the proof of Theorem 1.2, consider the map

$$\begin{aligned}\Gamma(u)=W(t)u_0&+\int_0^t W(t-\tau)(G(\Phi(u_0,u_1+\omega(T,u))))(\tau)d\tau\\ &+i\lambda\int_0^t W(t-\tau)\left(|u|^2u\right)(\tau)d\tau\end{aligned}$$

for any

$$u \in S_{M,o} = \Big\{ v \in X_T^{b,s};\ v \text{ is odd with respect to } x\text{-variable, } \|v\|_{X_T^{b,s}} \le M \Big\}$$

where, we recall that

$$\omega(T,u) := i\lambda \int_0^T W(T-\tau)\left(|u|^2 u\right)(\tau)d\tau.$$

Note that $\omega(T,u)$ is an odd function of x if u is an odd function of x. Thus, by Corollary 2.2, $G(\Phi(u_0, u_1 + \omega(T,u)))$ is an odd function of x, and consequently, $\Gamma(u)$ is an odd function of x for any $u \in S_{M,o}$. Then the same argument as in the proof of Theorem 1.1 shows that Γ has a fixed point in the set $S_{M,o}$ as long as $\|u_0\|_s + \|u_1\|_s$ is small enough and M is chosen accordingly. The proof is complete. □

Now we consider the system described by the nonlinear Schrödinger equation posed on the finite interval $(0,\pi)$ with the Neumann boundary conditions:

$$\text{(3.12)} \qquad \begin{cases} iw_t + w_{xx} + \lambda|w|^2 w = Qh_1, & x \in (0,\pi),\ t \ge 0, \\ w(x,0) = w_0(x), \qquad w_x(0,t) = 0, & w_x(\pi,t) = 0. \end{cases}$$

It has been shown in [1] that for given $s \ge 0$ and $w_0 \in H^s_{even}(0,\pi)$ and $h_1 \in L^1_{loc}(\mathbb{R}; H^s(0,\pi))$, (3.12) admits a unique solution $v \in C(\mathbb{R}; H^s(0,\pi))$. Moreover, if we let g be the even extension of q from the interval $(0,\pi)$ to the interval $(-\pi,\pi)$, and u_0 be the even extension of w_0 from the interval $(0,\pi)$ to the interval $(-\pi,\pi)$, if $u = u(x,t)$ denotes the even extension solution of $w(x,t)$ from the interval $(0,\pi)$ to the interval $(-\pi,\pi)$ with respect to the x-variable, then $u \in C(\mathbb{R}; H^s_p(-\pi,\pi))$ and solves (3.1) with h being the even extension of h_1 from the interval $(0,\pi)$ to the interval $(-\pi,\pi)$ with respect to the x-variable. On the other hand, if g is an even function, $u_0 \in H^s_p(-\pi,\pi)$ is an even function and h is also an even function with respect to the x-variable in (3.1), then the corresponding solution u of (3.1) is also an even function. If we let $w(x,t)$ be its restriction on the interval $(0,\pi)$, then v solves (3.12) with w_0, q, and h_1 being the restrictions of u_0, g and h to the interval $(0,\pi)$, respectively. This leads us to the proof of Theorem 1.3.

Proof of Theorem 1.3. It is exactly the same as the one of Theorem 1.2, except that all the odd extensions become even extensions. The proof is complete. □

Finally, we consider the boundary control of the nonlinear Schrödinger equation posed on the interval $(0,\pi)$ with the Dirichlet boundary conditions:

$$\text{(3.13)} \qquad \begin{cases} iu_t + u_{xx} + \lambda|u|^2 u = 0, & x \in (0,\pi),\ t \ge 0, \\ u(x,0) = u_0(x), \qquad v(0,t) = h(t), & u(\pi,t) = 0 \end{cases}$$

or with the Neumann boundary conditions:

$$\text{(3.14)} \qquad \begin{cases} iw_t + w_{xx} + \lambda|w|^2 w = 0, & x \in (0,\pi),\ t \ge 0, \\ w(x,0) = w_0(x), \qquad w_x(0,t) = h(t), & w_x(\pi,t) = 0. \end{cases}$$

Proof of Theorem 1.4. We prove only part (a). The proof of part (b) is similar. Consider the nonlinear Schrödinger equation posed on the finite interval $(-\pi,\pi)$ with Dirichlet boundary conditions:

$$\text{(3.15)} \qquad \begin{cases} iw_t + w_{xx} + \lambda|w|^2 w = ig(x)\mu(x,t), & x \in (-\pi,\pi),\ t \in \mathbb{R}, \\ w(x,0) = \tilde{u}_0(x), \quad x \in (-\pi,\pi), \\ w(-\pi,t) = 0, \qquad w(\pi,t) = 0, \quad t \in \mathbb{R}. \end{cases}$$

where g is supported in the interval $(-\pi,0)$ and $\mu(x,t)$ is a control input, and $\tilde{u}_0$ is an extension of $u_0 \in H^s(0,\pi)$ to the space $H^s(-\pi,\pi)$ satisfying $\tilde{u}_0(-\pi)=0$. For given $u_1 \in H^s(0,\pi)$ let $\tilde{u}_1$ be its extension to the space $H^s(-\pi,\pi)$. According to Theorem 1.2, one can find $\mu \in L^2(0,T;H^s(-\pi,\pi))$ such that (3.15) admits a unique solution $w \in C([0,T];H^s(-\pi,\pi))$ such that

$$w(x,0)=\tilde{u}_0(x), \qquad w(x,T)=\tilde{u_1}(x).$$

Let $u=u(x,t)$ be the restriction of $w(x,t)$ to the interval $(0,\pi)$. Then $u \in C([0,T]; H^s(0,\pi))$ solves (3.13) with $h(t):=w(0,t)$, and satisfies

$$u(x,0)=u_0(x), \qquad u(x,T)=u_1(x), \qquad x\in(0,\pi).$$

The proof is complete. □

4. Stabilization. In this section we study long-time behavior of the closed-loop system

$$\begin{cases} iu_t+u_{xx}+\lambda|u|^2u=-ig^2u, & x\in(-\pi,\pi),\ t\in R,\\ (u(x,0)=u_0(x), & x\in(-\pi,\pi),\\ u(-\pi,t)=u(\pi,t), & u_x(-\pi,t)=u_x(\pi,t). \end{cases} \tag{4.1}$$

We first consider the associated linear system:

$$\begin{cases} iu_t+u_{xx}=-ig^2u, & x\in(-\pi,\pi),\ t\in R,\\ u(x,0)=u_0(x), & x\in(-\pi,\pi),\\ u(-\pi,t)=u(\pi,t), & u_x(-\pi,t)=u_x(\pi,t). \end{cases} \tag{4.2}$$

For given $s\ge 0$, define an operator A_g in the space $H^s(-\pi,\pi)$ by

$$A_g=i\partial_x^2-g^2$$

with domain $\mathcal{D}(A_g)=H_p^{s+2}(-\pi,\pi)$. By the standard semigroup theory, it generates a continuous group $(W_g(t))_{t\in\mathbb{R}}$ of operators on $\mathcal{H}:=H_p^s(-\pi,\pi)$ and for any given initial data $u_0\in H^s(-\pi,\pi)$, the corresponding solution u of (4.2) can be expressed as

$$u(t)=W_g(t)u_0.$$

Moreover, the semigroup $(W_g(t))_{t\in\mathbb{R}^+}$ is exponentially stable in $\mathcal{H}$.

PROPOSITION 4.1. *There exist positive constants $C>0$ and $\nu>0$ such that*

$$||W_g(t)u_0||_s \le Ce^{-\nu t}||u_0||_s \qquad \forall t\ge 0. \tag{4.3}$$

Proof. When $s=0$, $g^2u=GG^*u$ and, thus, the exponential stability of $(W_g(t))_{t\in\mathbb{R}^+}$ is a direct consequence of Theorem 2.1 according to [12]. To prove (4.3) when $s=2$, we pick $u_0\in H_p^2(-\pi,\pi)$ and set $v:=u_t$. Then v solves the system

$$\begin{cases} v_t=iv_{xx}-g^2(x)v, & v(x,0)=v_0(x):=i\frac{d^2u_0}{dx^2}-g^2(x)u_0(x),\\ v(-\pi,t)=v(\pi,t), & v_x(-\pi,t)=v_x(\pi,t). \end{cases} \tag{4.4}$$

By the property (4.3) established when $s=0$, we have

$$||u(t)||_0\le Ce^{-\nu t}||u_0||_0, \qquad ||v(t)||_0\le Ce^{-\nu t}||v_0||_0.$$

Since $iu_{xx} = v + g^2u$, we conclude that

$$||u(t)||_2 \leq Ce^{-\nu t}||u_0||_2 \qquad \forall t \geq 0.$$

An easy induction yields (4.3) for any $s \in 2\mathbb{N}$. The proposition follows by a classical interpolation argument. The proof is complete. $\square$

Now we turn our attention to the stability properties of the nonlinear system (4.1) which can be rewritten in equivalent its integral form

$$u(t) = W_g(t)u_0 + i\lambda \int_0^t W_g(t-\tau)\left(|u|^2u\right)(\tau)d\tau. \tag{4.5}$$

At this point, we need to establish Lemmas 3.1 and 3.2 with the semigroup $W(t)$ substituted by the semigroup $W_g(t)$.

Lemma 4.1. *Let $T > 0$, $s \geq 0$, and $0 \leq b \leq 1$ be given. There exists a constant $C > 0$ depending only on s and b such that*

$$\|W_g(t)\phi\|_{X_T^{b,s}} \leq C\|\phi\|_s$$

for any $\phi \in H_p^s(-\pi, \pi)$.

Proof. An application of Duhamel formula gives

$$W_g(t)\phi = W(t)\phi - \int_0^t W(t-\tau)\left(g^2W_g(\tau)\phi\right)d\tau. \tag{4.6}$$

It follows that

$$\begin{aligned}||W_g(t)\phi||_{X_T^{b,s}} &\leq ||W(t)\phi||_{X_T^{b,s}} + \left\|\int_0^t W(t-\tau)\left(g^2W_g(\tau)\phi\right)d\tau\right\|_{X_T^{b,s}} \\ &\leq C||\phi||_s + C\left\|g^2W_g(t)\phi\right\|_{X_T^{b'-1,s}} \\ &\leq C||\phi||_s + C||W_g(t)\phi||_{L^2(0,T;H_p^s(-\pi,\pi))} \qquad (\text{as } b'-1<0) \\ &\leq C||\phi||_s,\end{aligned}$$

as desired. $\square$

Lemma 4.2. *Let $T > 0$, $s \geq 0$, $b \in (\frac{1}{2}, \frac{5}{8})$, and $b' \in (b, \frac{5}{8})$ be given. There exists a constant $C > 0$ depending only on s, b, and b' such that*

$$\left\|\int_0^t W_g(t-\tau)f(\tau)d\tau\right\|_{X_T^{b,s}} \leq C\|f\|_{X_T^{b'-1,s}}$$

for any $f \in X_T^{b'-1,s}$.

Proof. It follows from (4.6) that

$$\int_0^t W_g(t-\tau)f(\tau)d\tau = \int_0^t W(t-\tau)f(\tau)d\tau - \int_0^t W(t-\tau)g^2\left(\int_0^\tau W_g(\tau-s)f(s)ds\right)d\tau,$$

and hence, using Lemma 3.2,

$$\begin{aligned}\left\|\int_0^t W_g(t-\tau)f(\tau)d\tau\right\|_{X_T^{b,s}} &\leq C||f||_{X_T^{b'-1,s}} + C\left\|g^2\int_0^t W_g(t-s)f(s)ds\right\|_{X_T^{b'-1,s}} \\ &\leq C||f||_{X_T^{b'-1,s}} + CT^\alpha||g||^2_{X_T^{b,s}}\left\|\int_0^t W_g(t-s)f(s)ds\right\|_{X_T^{b,s}}\end{aligned}$$

for some constant $\alpha > 0$, by virtue of Lemma 3.3. The result follows at once if T is small enough, say $T < T_0$. For $T \geq T_0$, the result follows from Lemma 4.1 and an easy induction argument. □

Now we are in position to prove Theorem 1.5

Proof of Theorem 1.5. For given $s \geq 0$, there exist some positive constants C, ν such that

$$||W_g(t)u_0||_s \leq Ce^{-\nu t}||u_0||_s \qquad \forall t \geq 0$$

according to Proposition 4.1. Choose $T > 0$ such that

$$Ce^{-\nu T} < \frac{1}{4}$$

and fix a number $b \in (\frac{1}{2}, \frac{5}{4})$. We seek a solution u to the integral equation (4.5) as a fixed point of the map

$$\Gamma(u) = W_g(t)u_0 + i\lambda \int_0^t W_g(t-\tau)\left(|u|^2 u\right)(\tau)d\tau$$

in some ball S_M of the space $X_T^{b,s}$. This will be done provided that $||u_0||_s \leq \delta$ where δ is a small number to be determined. Furthermore, to ensure the exponential stability, δ and M will be chosen in such a way that $||u(T)||_s \leq ||u_0||_s/2$. Pick for the moment any $\delta > 0$ and $M > 0$, and let $u_0 \in \mathcal{H}$ be such that $||u_0||_s \leq \delta$. By computations similar to the ones displayed in the proof of Theorem 1.1 with $W_g(t)$ substituted to $W(t)$, we arrive to

$$||\Gamma(u)||_{X_T^{b,s}} \leq c||u_0||_s + cM^3 \qquad \forall u \in S_M$$

and

$$||\Gamma(u) - \Gamma(v)||_{X_T^{b,s}} \leq cM^2||u-v||_{X_T^{b,s}}$$

for some constant $c > 0$ independent of δ, M, and t. On the other hand, using the estimate of $||\omega(T,u)||_s$ in the proof of Theorem 1.1, we obtain

$$\begin{aligned} ||\Gamma(u)(T)||_s &\leq ||W_g(T)u_0||_s + \left\| \int_0^T W_g(T-t)(|u|^2u)(t)dt \right\|_s \\ &\leq \frac{1}{4}||u_0||_s + cM^3. \end{aligned}$$

Pick $\delta = 4cM^3$ where $M > 0$ is chosen so that

$$\left(4c^2 + c\right) M^3 \leq M \;\text{ and }\; cM^2 \leq \frac{1}{2}.$$

Then we have

$$\begin{aligned} ||\Gamma(u)||_{X_T^{b,s}} &\leq M \qquad \forall u \in S_M \\ ||\Gamma(u) - \Gamma(v)||_{X_T^{b,s}} &\leq \frac{1}{2}||u-v||_{X_T^{b,s}} \qquad \forall u, v \in S_M. \end{aligned}$$

Therefore, Γ is a contraction in S_M. Furthermore, its unique fixed point $u \in S_M$ fulfills

$$||u(T)||_s = ||\Gamma(u)(T)||_s \leq \frac{\delta}{2}.$$

Assume now that $0 < ||u_0||_s < \delta$. Changing δ into $\delta' := ||u_0||_s$ and M into $M' := (\delta'/\delta)^{\frac{1}{3}}M$, we infer that $||u(T)||_s \leq ||u_0||_s/2$, and an obvious induction yields $||u(nT)||_s \leq 2^{-n}||u_0||_s$ for any $n \geq 0$. As $X_T^{b,s} \subset C([0,T]; H_p^s(-\pi,\pi))$ for $b > 1/2$, and $||u||_{X_T^{b,s}} \leq M = (\delta/(4c))^{\frac{1}{3}}$, we infer by the semigroup property that there exist some constants $C' > 0, \nu' > 0$ such that

$$||u(t)||_s \leq C' e^{-\nu' t}||u_0||_s.$$

The proof is complete. □

REFERENCES

[1] J. L. Bona, S. M. Sun, and B.-Y. Zhang, *Nonhomogeneous boundary value problems of the nonlinear Schrödinger equation*, in preparation.

[2] J. Bourgain, *Fourier transform restriction phenomena for certain lattice subsets and applications to non-linear evolution equations, part* I: *Schrödinger equations*, Geom. Funct. Anal., 3 (1993), pp. 107–156.

[3] J. Bourgain, *Global Solutions of Nonlinear Schrödinger Equations*, Colloqium Publication, Vol. 46, American Mathematical Society, Providence, RI, 1999.

[4] N. Burq and M. Zworski, *Geometric control in the presence of a black box*, J. Am. Math. Soc., 17 (2004), pp. 443–471.

[5] B. Dehman, P. Gérard, and G. Lebeau, *Stabilization and control for the nonlinear equation on a compact surface*, Math. Z, 254 (2006), pp. 729–749.

[6] R. Illner, H. Lange, and H. Teismann, *Limitations on the control of Schrödinger equations*, ESAIM Control Optim. Calc. Var., 12 (2006), pp. 615–635.

[7] R. Illner, H. Lange, and H. and Teismann, *A note on the exact internal control of nonlinear Schrödinger equations*, CRM Proc. Lecture Notes, 33 (2003), pp. 127–137.

[8] S. Jaffard, *Contrôle interne des vibrations d'une plaque rectangulaire*, Port. Math., 47 (1990), pp. 423–429.

[9] H. Lange and H. Teismann, *Controllability of nonlinear Schrödinger equation in the vicinity of the ground state*, Math. Methods Appl. Sci., 30 (2007), pp. 1483–1505.

[10] I. Lasiecka and R. Triggiani, *Optimal regularity, exact controllability and uniform stabilization of Schrödinger equaitons with Dirichlet control,* Differ. Integral Equ., 5 (1992), pp. 521–535.

[11] G. Lebeau, *Contrôle de l'équation de Schrödinger*, J. Math. Pures Appl., 71 (1992), pp. 267–792.

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

[13] E. Machtyngier, *Contrôlabilité exacte et stabilisation frontière de l'équation de Schrödinger*, C. R. Acad. Sci. Paris, 310 (1990), pp. 801–806.

[14] E. Machtyngier, *Exact controllability for the Schrödinger equation,* SIAM J. Control Optim., 32 (1994), pp. 24–34.

[15] E. Machtyngier and E. Zuazua, *Stabilization of the Schrödinger equation*, Port. Math., 51 (1994), pp. 243–256.

[16] A. Mercado, A. Osses, and L. Rosier, *Inverse problems for the Schrödinger equation via Carleman inequalities with degenerate weights*, Inverse Problems, 24 (2008), 015017.

[17] L. Miller, *How violent are fast controls for Schrödinger and plate vibrations?* Arch. Ration. Mech. Anal., 172 (2004), pp. 429–456.

[18] A. Pazoto and L. Rosier, *Stabilization of a Boussinesq system of KdV-KdV type*, Systems & Control Lett., 57 (2008), pp. 595–601.

[19] G. Perla Menzala, C. F. Vasconcellos, and E. Zuazua, *Stabilization of the Korteweg-de Vries equation with localized damping*, Quart. Appl. Math. 60 (2002), pp. 111–129.

[20] K.-D. Phung, *Observability and controllability for Schrödinger equations*, SIAM J. Control Optim., 40 (2001), pp. 211–230.

[21] K. Ramdani, T. Takahashi, G. Tenenbaum, and M. Tucsnak, *A spectral approach for the exact observability of infinite-dimensional systems with skew-adjoint generator*, J. Funct. Anal., 226 (2005) pp. 193–229.

[22] L. Rosier, *Exact boundary controllability for the Korteweg-de Vries equation on a bounded domain*, ESAIM: Control Optim. Calc. Var., 2 (1997), pp. 33–55.

[23] L. Rosier and B.-Y. Zhang, *Global stabilization of the generalized Korteweg-de Vries equation*, SIAM J. Control Optim., 45 (2006), pp. 927–956.

[24] D. L. Russell and B.-Y. Zhang, *Controllability and stabilizability of the third order linear dispersion equation on a periodic domain*, SIAM J. Control Optim., 31 (1993), pp. 659–676.

[25] D. L. Russell and B.-Y. Zhang, *Exact controllability and stabilizability of the Korteweg-de Vries equation*, Trans. Amer. Math. Soc., 348 (1996), pp. 3643–3672.

[26] R. M. Young, *An Introduction to Nonharmonic Fourier Series*, Academic Press, New York, 1980.

[27] V. E. Zakharov, *Stability of periodic waves of finite amplitude on the surface of a deep fluid*, J. Appl. Mech. Tech. Phys., 9 (1968), pp. 190–194.

[28] B.-Y. Zhang, *Exact boundary controllability of the Korteweg-de Vries equation*, SIAM J. Control Optim., 37 (1999), pp. 543–565.

[29] E. Zuazua, *Remarks on the controllability of the Schrödinger equation*, Quantum control: Mathematical and numerical challenges, CRM Proc. Lecture Notes, 33, Amer. Math. Soc., Providence, RI, 2003, pp. 193–211.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 2, pp. 993–1030

© 2009 Society for Industrial and Applied Mathematics

ON A CLASS OF OPTIMIZATION PROBLEMS FOR FINITE TIME HORIZON INVENTORY MODELS*

LAKDERE BENKHEROUF† AND BRIAN H. GILDING‡

Abstract. The paper proposes a general theory for the treatment of a class of optimization problems that arise as a consequence of a search for the optimal replenishment schedule for finite time horizon inventory models. The decision variables in these optimization problems consist of the number of replenishment periods and the lengths of the periods. When the number of replenishment periods is fixed, it is shown that the optimization problems have a unique solution under some partial differential inequalities. Furthermore, the minimal value is convex as a function of the number of replenishment periods. This leads to a simple procedure for determining the optimal control policy.

Key words. inventory model, finite time horizon, optimal schedule, deteriorating item

AMS subject classification. 90B05

DOI. 10.1137/070683945

1. Introduction.

This paper examines the problem of finding the optimal replenishment schedule for a family of generalized inventory models where the planning horizon is finite. These models include the classical models with stock- and time-dependent demand item, a variable deterioration rate, as well as shortages [2, 3, 5, 6, 9, 12]. Related models are treated in [14, 17, 18]. In the papers just cited, it is shown that for a fixed number of replenishment periods, n say, the first-order condition for optimality (that is, the solution to the system of nonlinear equations obtained from setting the first partial derivatives equal to zero) is enough to characterize the optimal replenishment schedule uniquely. Moreover, the search for the solution of the system of nonlinear equations generated by the first-order condition can be reduced to a univariate line search method. Proofs of optimality are based on showing that the Hessian matrix associated with the optimization problem is positive definite [2, 4, 9, 12, 14]. The convexity of the optimal cost function of the inventory system with respect to n [2, 14, 17, 18] provides the justification for algorithms to find the optimal replenishment schedule.

Notwithstanding the progress outlined above, the search for a general class of inventory models for which there is an optimal replenishment schedule remains ongoing. This paper, we hope, is a step forward toward resolving this search.

To motivate the presentation in this article, we start by considering models in which no shortages are assumed. Models with shortages will be treated in subsequent sections. The common assumptions behind existing models neglecting shortages [2, 5, 11, 15] are the following:

1. A single item is held in stock over a known and finite planning horizon $0 \leq t \leq H$, where t denotes time and $H > 0$.
2. The level of stock on hand $I(t)$ is depleted by the combined effect of demand

*Received by the editors February 28, 2007; accepted for publication (in revised form) July 8, 2008; published electronically March 6, 2009. This work was supported by SQU fund IG/SCI/DOMS/03/02.

http://www.siam.org/journals/sicon/48-2/68394.html

†Department of Statistics and Operations Research, College of Science, Kuwait University, P.O. Box 5969, Safat 13060, Kuwait (lakdereb@yahoo.com).

‡Department of Mathematics and Statistics, College of Science, Sultan Qaboos University, P.O. Box 36, Al-Khodh 123, Oman (gilding@squ.edu.om).

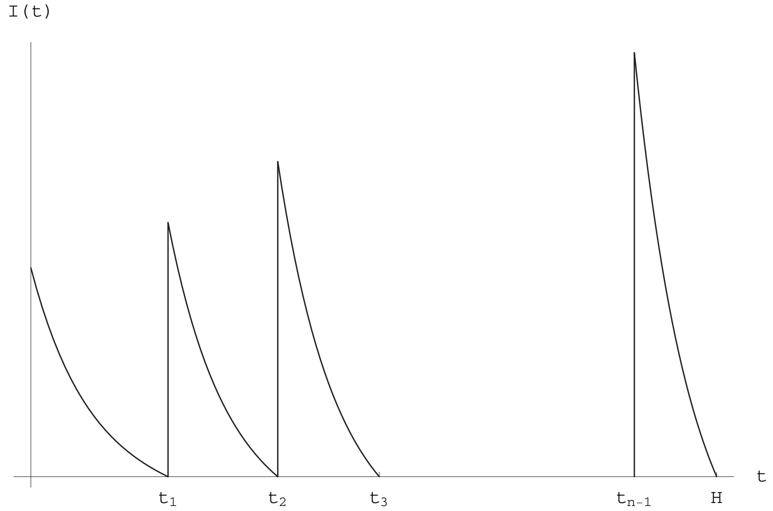


FIG. 1. *Typical inventory behavior.*

and deterioration.

3. Deteriorated units are neither repaired nor replaced.
4. Replenishment occurs instantaneously (at an infinite rate) when the inventory level decreases to zero.
5. The cost structure is
   (a) a fixed setup cost K,
   (b) a holding cost per unit in stock per unit of time c_1,
   (c) the inventory holding cost is charged only for good units,
   (d) a purchasing cost c_2.

Figure 1 shows a typical behavior of the inventory during the planning horizon H, where n denotes the number of replenishment periods and t_j denotes the time elapsed at the end of the jth period. For completeness, we define $t_0 = 0$ so that for $j = 1, 2, \ldots, n$, the jth replenishment period is the time interval (t_{j-1}, t_j), with

$$0 = t_0 < t_1 < t_2 < \cdots < t_{n-1} < t_n = H. \tag{1.1}$$

The standard model of the dynamics of the inventory level during a replenishment period is

$$I'(t) = -D(t) - \theta(t)I(t), \tag{1.2}$$

in which D is a continuously differentiable monotonic positive function of time corresponding to the demand rate and θ is a continuously differentiable monotonic nonnegative function of time denoting the rate of deterioration [3, 6, 9, 10, 11, 15]. Here, and throughout the remainder of the paper, we use the following notation.

$'$: The derivative of a univariate function.
∂_x : The partial derivative of a bivariate function with respect to the first variable.
∂_y : The partial derivative of a bivariate function with respect to the second variable.

Several products are known to experience variation in their demand. The demand for certain products may be seasonal, for instance, that for warm clothes or oil which is high in winter. Also, the age of the inventory may have a negative impact on the demand due to loss of quality. Deterioration of the product refers to spoilage, damage dryness, vaporization, etc. Products like food stuffs, medicine, and blood are known to perish over time. Moreover, certain products experience a different rate of deterioration with time, for instance, some vegetables perish faster in summer than winter. Starting from the assumption that both of the functions D and θ are positive constants [10], the model (1.2) has evolved over a number of years based on the observation of inventory systems and its mathematical tractability.

An obvious drawback of (1.2) is that it ignores stock-dependent demand. It is a common phenomena in supermarkets that large piles of displayed goods attract customers. Hence, a retailer may influence the demand by displaying large of quantities of goods. As a result, there is a need to examine inventory models with time-varying stock-dependent demand items and time-varying deterioration. For more details, see [7, 11, 15, 16] and the references therein.

To date, the dynamics of an inventory with stock-dependent demand have been modelled almost exclusively by the equation

$$I'(t) = -G(t)I^{\beta}(t) - \theta I(t), \tag{1.3}$$

where G is a continuously differentiable positive function of time, β is a constant within the range $0 < \beta < 1$, and θ is a nonnegative constant. With G a constant function and $\theta = 0$, this model was proposed in [1], extended to arbitrary θ in [13], and enlarged to the above form in [2], where it was studied in depth for $0 \leq \beta < 1$. The model is a realistic and logical extension of (1.2), to which it reduces when $\beta = 0$. The parameter β may be interpreted as the elasticity of the demand with respect to the inventory level. The model itself is amenable to study because of the power-law dependence of the demand on the level of stock.

A variation on (1.3) with G constant and $\theta = 0$ was proposed in [8]. This model takes the form

$$I'(t) = -G\left(\max\{I(t), I_0\}\right)^{\beta}, \tag{1.4}$$

with I_0 a positive constant. The idea behind this model is that the demand is stock-dependent when the inventory level exceeds a critical level I_0. When the inventory lies below this critical level, the demand is constant.

It is clear though that few inventory systems are accurately described by such idealized dynamics as those embedded in (1.2)–(1.4). The generalization of (1.2)–(1.4) that we wish to consider is

$$I'(t) = -D(t)F(e^{\Theta(t)}I(t)) - \theta(t)I(t), \tag{1.5}$$

in which $D(t)$ is positive and $\theta(t)$ is nonnegative for $0 < t < H$ and

$$\Theta(t) = \int_0^t \theta(u)\,du.$$

With regard to smoothness, the functions D and θ are assumed to be $C([0,H]) \cap C^1(0,H)$, where $C([0,H])$ denotes the class of continuous functions on the closed interval $[0,H]$ and $C^1(0,H)$ denotes the class of continuously differentiable functions on the open interval $(0,H)$. Letting

$$\mu(x,y) = \int_x^y D(t)\,e^{\Theta(t)}\,dt,$$

it is assumed that $F(v)$ is positive for $0 < v < \ell$ and $F \in C(0, \ell)$, for some number ℓ for which

$$\mu(0, H) \le \int_0^\ell \frac{dv}{F(v)} < \infty. \tag{1.6}$$

The standard model (1.2) is clearly a special case of (1.5) with $F \equiv 1$. Moreover, even without the restrictions $\beta > 0$ and θ is constant, (1.3) is a special case of (1.5) with $D(t) = G(t)\, e^{-\beta\Theta(t)}$ and $F(v) = v^\beta$. Model (1.4) can be cast in the form (1.5) by taking $D \equiv 1$ and $F(v) = G(\max\{v, I_0\})^\beta$. Generalization (1.5) is motivated by the desire to increase the scope of describing practical inventory systems while retaining mathematical tractability. A concrete example of (1.5) that has not been treated before is provided by $F(v) = v\,|\ln v|^\alpha$ for some constant $\alpha > 1$. This leads to

$$I'(t) = -D(t)\, e^{\Theta(t)}\, |\ln I(t) + \Theta(t)|^\alpha\, I(t) - \theta(t) I(t). \tag{1.7}$$

Equation (1.5) holds for $t_{j-1} \le t < t_j$, with the boundary condition

$$I(t) \to 0 \quad \text{as } t \uparrow t_j \tag{1.8}$$

for $j = 1, 2, \ldots, n$. To solve the equation subject to this condition, we change the dependent variable to

$$J(t) = e^{\Theta(t)} I(t).$$

With this as the unknown, (1.5) becomes

$$J'(t) = -D(t)\, e^{\Theta(t)} F(J(t)).$$

Subsequently, by separation of variables and imposition of (1.8), we find

$$\int_0^{J(t)} \frac{dv}{F(v)} = \mu(t, t_j)$$

for $t_{j-1} \le t < t_j$. Hence, if we define ψ via

$$z = \int_0^{\psi(z)} \frac{dv}{F(v)} \quad \text{for } 0 \le z \le \mu(0, H), \tag{1.9}$$

we have

$$J(t) = \psi(\mu(t, t_j)) \quad \text{for } t_{j-1} \le t < t_j.$$

This gives

$$I(t) = \kappa(t, t_j) \quad \text{for } t_{j-1} \le t < t_j, \tag{1.10}$$

where

$$\kappa(x, y) = e^{-\Theta(x)} \psi(\mu(x, y)). \tag{1.11}$$

For classical model (1.2), we find (1.10), (1.11) with

$$\psi(z) = z.$$

For the case of (1.3), we have (1.10), (1.11) with

$$\psi(z) = \{(1-\beta)z\}^{1/(1-\beta)} \tag{1.12}$$

and $D(t) = G(t)\, e^{-\beta\Theta(t)}$. Model (1.4) leads to

$$\psi(z) = \begin{cases} G I_0^{\beta} z & \text{for } z \le I_0^{1-\beta}/G, \\ \left\{\beta I_0^{1-\beta} + (1-\beta)Gz\right\}^{1/(1-\beta)} & \text{for } z > I_0^{1-\beta}/G. \end{cases}$$

In the case of (1.7), we obtain (1.10), (1.11) with

$$\psi(z) = \exp\left(-\{(\alpha-1)z\}^{-1/(\alpha-1)}\right).$$

Note that if $F(0) = 0$, the solution of (1.5) satisfying (1.8) is not unique. There exists a one parameter family of solutions given by

$$I(t) = \begin{cases} \kappa(t,\eta) & \text{for } t_{j-1} \le t < \eta, \\ 0 & \text{for } \eta \le t < t_j, \end{cases}$$

and $\eta \in [t_{j-1}, t_j]$. This is true even in the case of (1.3) with $0 < \beta < 1$. For our purpose, we shall adopt solution (1.10). This is the only member of the family that conforms to modeling assumption 4.

When the evolution of the level of stock is as described above, one may adopt two different functions for the total cost $\mathcal{C}$. The first is the sum of the costs of ordering, holding, and deterioration (OHD). For a fixed number of replenishment periods n, this is given by

$$\mathcal{C} = nK + \sum_{j=1}^{n} \left\{ c_1 \int_{t_{j-1}}^{t_j} I(t)\, dt + c_2 \int_{t_{j-1}}^{t_j} \theta(t) I(t)\, dt \right\}. \tag{1.13}$$

Thus, the total cost

$$\mathcal{C} = nK + \sum_{j=1}^{n} \left\{ \int_{t_{j-1}}^{t_j} w(t)\kappa(t,t_j)\, dt \right\}, \tag{1.14}$$

where

$$w(t) = c_1 + c_2\theta(t). \tag{1.15}$$

The alternative cost function is the sum of the costs of ordering, holding, and purchasing (OHP). This is

$$\begin{aligned} \mathcal{C} &= nK + \sum_{j=1}^{n} \left\{ c_1 \int_{t_{j-1}}^{t_j} I(t)\, dt + c_2 I(t_{j-1}) \right\} \\ &= nK + \sum_{j=1}^{n} \left\{ \int_{t_{j-1}}^{t_j} c_1\kappa(t,t_j)\, dt + c_2\kappa(t_{j-1},t_j) \right\}. \end{aligned} \tag{1.16}$$

Henceforth, we shall refer to (1.14) and (1.16) as the OHD and OHP cost functions, respectively. The results of the paper apply to both.

Finding the optimal replenishment schedule reduces to the problem of finding n and $t_0, t_1, \ldots, t_n$ which minimizes $\mathcal{C}$ subject to constraint (1.1). When the dynamics of the inventory are described by linear equation (1.2), it can be determined that the OHP cost function differs from the OHD cost function merely by the addition of a fixed amount $c_2 \int_0^H D(t)\,dt$. Thus, in the classical case, the OHD and OHP optimization problems are equivalent. However, when the inventory dynamics are described by another equation, such as (1.3) with $\beta \neq 0$ or (1.5) with $F \not\equiv 1$, the OHD and OHP optimization problems are intrinsically distinct.

In [2] the optimal replenishment schedule was obtained for the OHP cost function (1.16) based on model (1.3) with $0 \leq \beta < 1$. Let us briefly outline the approach. Assume that n is known and ignore constraint (1.1) with the exception of the assumption that $t_n = H$. In this case, the first-order condition for optimality yields

$$\int_{t_{j-1}}^{t_j} c_1 (\partial_y \kappa)(t, t_j)\,dt + c_2 (\partial_y \kappa)(t_{j-1}, t_j) = (c_1 \kappa - c_2 \partial_x \kappa)(t_j, t_{j+1}) \tag{1.17}$$

for $j = 1, 2, \ldots, n-1$. Now, if one selects $t_{n-1} < H$, then t_{n-2} can be obtained uniquely, since the left-hand side of (1.17) is strictly decreasing in t_{j-1}. This iterative process can then be repeated to find t_{n-3} and so on, down to t_0. Automatically, $t_0 < t_1 < \cdots < t_{n-1} < t_n = H$. Using the notation $t_0 = t_0(t_{n-1})$ to emphasize the dependence of t_0 on t_{n-1}, the problem of solving (1.17) subject to (1.1) reduces to that of finding t_{n-1} such that $t_0(t_{n-1}) = 0$. It was shown that a unique solution based on the above iterative process can be found when the dynamics of the inventory is given by (1.3) with θ constant, G' positive, and G'/G nonincreasing. This solution was proven to be optimal by an examination of the Hessian matrix of the optimization problem. The optimal value of the cost function could be shown to be convex in n following the approach in [17].

The cost function associated with either the OHD model (1.14) or the OHP model (1.16) may be written as

$$\mathcal{C} = nK + S_n(t_0, t_1, \ldots, t_n), \tag{1.18}$$

where

$$S_n(t_0, t_1, \ldots, t_n) = \sum_{j=1}^{n} R(t_{j-1}, t_j) \tag{1.19}$$

and R is a real-valued function with domain

$$\Omega = \{(x, y) : 0 \leq x < y \leq H\}.$$

For OHD model (1.14), we have

$$R(x, y) = \int_x^y w(t) \kappa(t, y)\,dt, \tag{1.20}$$

with w given by (1.15) and for OHP model (1.16),

$$R(x, y) = \int_x^y c_1 \kappa(t, y)\,dt + c_2 \kappa(x, y). \tag{1.21}$$

Guided by prior results, the next section contains theoretical material on the optimality of cost functions pertaining to the models presented above and to further models where shortages are allowed. It is shown that a unique optimal solution exists under some partial differential inequalities. These inequalities are satisfied by the existing models in the literature and other models not previously treated.

The analysis of this paper is new in nature and avoids demonstrating that the Hessian matrix of the cost function is positive definite. The latter has been previously employed to show that the solution obtained from the first-order optimality conditions is indeed the optimal solution, For all but the simplest of models, it involves lengthy and arduous computations [2, 4, 9, 12, 14]. The present approach yields the optimality of the solution as part and parcel of the proof of existence and circumvents messy calculations. Consequentially, it reveals that certain assumptions that have been imposed in the past may be weakened or even discarded. Moreover, the subsequent proof of convexity is cleaner and more straightforward.

Although the optimization problems treated in this paper have been developed for a particular application, namely, the search for an optimal replenishment schedule for finite time horizon inventory models, they could also be examined as a purely mathematical optimization problem.

The next section contains the main optimality result when n is fixed as well as a detailed description of the hypotheses under which this result is obtained. Section 3 presents conditions on specific inventory models for which the optimality result of section 2 applies. Section 4 deals with an extension of the optimality result to handle shortages. The last section contains some general conclusions.

2. The general model and optimality. All the models with or without shortages that we shall consider have the form

$$\mathcal{C} = \nu_n K + S_n(t_0, t_1, \dots, t_n), \tag{2.1}$$

where ν_n denotes the number of orders made,

$$S_n(t_0, t_1, \dots, t_n) = \sum_{j=1}^{n} R_j(t_{j-1}, t_j), \tag{2.2}$$

and $\{R_j\}_{j=1}^{n}$ is a set of functions defined on Ω, satisfying the following generic hypothesis. For clarity, $C^p(\Omega)$ denotes the set of real functions defined on the interior of Ω for which every partial derivative of order less than or equal to p exists and is continuously extendible to Ω, and $C^p(\overline{\Omega})$ denotes the subset for which these derivatives are continuously extendible to $\overline{\Omega}$, i.e., the closure of Ω.

Hypothesis 1. For every $j \geq 1$, the function $R_j \in C^1(\overline{\Omega}) \cap C^2(\Omega)$ is such that

$$R_j > 0 \quad \text{in } \Omega, \tag{2.3}$$

$$R_j = 0 \quad \text{on } \overline{\Omega} \setminus \Omega, \tag{2.4}$$

$$\partial_x R_j < 0 < \partial_y R_j \quad \text{in } \Omega, \tag{2.5}$$

and

$$\partial_x \partial_y R_j < 0 \quad \text{in } \Omega. \tag{2.6}$$

Note that (2.3) is in some sense redundant, since it can be deduced from (2.4) and (2.5).

For both of the models described in the previous section, $\nu_n = n$. The interpretation of ν_n for other models will be made more precise in section 4. For the OHD cost function, R_j is given by (1.20) for all j and for the OHP cost function, it is given by (1.21) for all j. At this stage, the subscript j could be dropped from the notation of R. However, it will be kept throughout this section as it will be found useful when dealing with models with shortages in section 4.

We shall initially assume that n and consequently ν_n is fixed, as in the treatment of [2], and concentrate on minimizing S_n. To do this, we require Hypothesis 1 and the following.

Hypothesis 2. There holds

$$\partial_y R_j + \partial_x R_{j+1} = 0 \quad \text{on } \overline{\Omega} \setminus \Omega \tag{2.7}$$

for all $1 \leq j \leq n-1$. Moreover, there is a function $f \in C(0,H)$ such that

$$\mathcal{L}_x R_{j+1} \geq 0 \quad \text{and} \quad \mathcal{L}_y R_j \geq 0 \quad \text{in } \Omega \tag{2.8}$$

for all $1 \leq j \leq n-1$, where

$$\mathcal{L}_x z = \partial_x^2 z + \partial_x \partial_y z + f(x)\partial_x z \tag{2.9}$$

and

$$\mathcal{L}_y z = \partial_x \partial_y z + \partial_y^2 z + f(y)\partial_y z. \tag{2.10}$$

Remark 1. The second part of Hypothesis 2 can be weakened to the assumption that there is a sequence of functions $\{f_j\}_{j=1}^{n-1} \subset C(0,H)$ such that $\mathcal{L}_x^{(j)} R_{j+1} \geq 0$ and $\mathcal{L}_y^{(j)} R_j \geq 0$ in Ω for all $1 \leq j \leq n-1$, where $\mathcal{L}_x^{(j)} z = \partial_x^2 z + \partial_x \partial_y z + f_j(x)\partial_x z$ and $\mathcal{L}_y^{(j)} z = \partial_x \partial_y z + \partial_y^2 z + f_j(y)\partial_y z$.

Under Hypotheses 1 and 2, our central result on optimality is the following.

THEOREM 1. *The function S_n given by* (2.2) *has a unique minimum with respect to $t_0, t_1, \ldots, t_n$ satisfying* (1.1).

The proof is by induction on the number of replenishment periods n. Moreover, it establishes the existence of a sequence of functions $\{\tau_j\}_{j=0}^{n-1} \subset C([0,H]) \cap C^1(0,H)$, with

$$\tau_j(0) = 0 \tag{2.11}$$

and

$$0 < \tau_j'(\eta) < 1 \tag{2.12}$$

for $0 < \eta < H$ and $j \geq 1$, such that, for any $0 < h \leq H$, the minimum of S_n under the constraint

$$0 = t_0 < t_1 < t_2 < \cdots < t_{n-1} < t_n = h \tag{2.13}$$

is given by $t_n = h$ and

$$t_j = \tau_j(t_{j+1}) \quad \text{for } j = n-1, n-2, \ldots, 0. \tag{2.14}$$

The induction hypothesis is the following.

Hypothesis 3. There exist functions $\{\tau_j\}_{j=0}^{N-1} \subset C([0,H]) \cap C^1(0,H)$ such that

$$\tau_0(\eta) = 0 \quad \text{for } 0 \le \eta \le H, \tag{2.15}$$

(2.11) and (2.12) hold for $0 < \eta < H$ and $1 \le j \le N-1$, and

$$(\partial_y R_j)(\tau_{j-1}(\tau_j(\eta)), \tau_j(\eta)) + (\partial_x R_{j+1})(\tau_j(\eta), \eta) = 0 \tag{2.16}$$

for all $0 < \eta < H$ and $1 \le j \le N-1$, with the following property. For every $0 < h \le H$ and $1 \le n \le N$, the function S_n given by (2.2) has a unique minimum with respect to $t_0, t_1, \dots, t_n$ satisfying (2.13), which is given by $t_n = h$ and (2.14). Furthermore, if $s_n(h)$ denotes the associated minimum value of S_n, then

$$s_n'(h) = (\partial_y R_n)(\tau_{n-1}(h), h) \tag{2.17}$$

for all $0 < h < H$ and $1 \le n \le N$.

Proof of Theorem 1. The induction hypothesis holds for $N = 1$, since under the constraint (2.13), $S_1(t_0, t_1) = S_1(0, h) = R_1(0, h)$.

Let us now assume that the induction hypothesis is true for a certain integer $N \ge 1$ and consider the problem of minimizing S_n with respect to $t_0, t_1, \dots, t_n$ subject to the constraint (2.13) for $n = N+1$. By Bellman's principle of optimality and the induction hypothesis, this problem is equivalent to that of minimizing $\sigma_{N+1}(t_N, h)$ subject to the constraint $0 < t_N < h$, where

$$\sigma_{N+1}(\eta, h) = s_N(\eta) + R_{N+1}(\eta, h) \tag{2.18}$$

and the remaining components of the minimum of S_n are given by $t_j = \tau_j(t_{j+1})$ for $j = N-1, N-2, \dots, 0$.

Using (2.17) and (2.18), we calculate that

$$(\partial_x \sigma_{N+1})(\eta, h) = (\partial_y R_N)(\tau_{N-1}(\eta), \eta) + (\partial_x R_{N+1})(\eta, h), \tag{2.19}$$

$$(\partial_y \sigma_{N+1})(\eta, h) = (\partial_y R_{N+1})(\eta, h), \tag{2.20}$$

$$(\partial_x \partial_y \sigma_{N+1})(\eta, h) = (\partial_x \partial_y R_{N+1})(\eta, h), \tag{2.21}$$

and

$$\begin{aligned}\left(\partial_x^2 \sigma_{N+1}\right)(\eta, h) = {} & (\partial_x \partial_y R_N)(\tau_{N-1}(\eta), \eta)\tau_{N-1}'(\eta) \\ & + \left(\partial_y^2 R_N\right)(\tau_{N-1}(\eta), \eta) + \left(\partial_x^2 R_{N+1}\right)(\eta, h).\end{aligned}$$

Employing (2.9) and (2.10), for later use, we rewrite the last expression as

$$\begin{aligned}\left(\partial_x^2 \sigma_{N+1}\right)(\eta, h) = {} & -f(\eta)(\partial_x \sigma_{N+1})(\eta, h) \\ & + (\mathcal{L}_x R_{N+1})(\eta, h) + (\mathcal{L}_y R_N)(\tau_{N-1}(\eta), \eta) \\ & - (\partial_x \partial_y R_N)(\tau_{N-1}(\eta), \eta)\{1 - \tau_{N-1}'(\eta)\} \\ & - (\partial_x \partial_y \sigma_{N+1})(\eta, h),\end{aligned} \tag{2.22}$$

where f is the function from Hypothesis 2.

Fix $0 < h \le H$. From (2.6), (2.7)–(2.12), and (2.19), it follows that

$$\begin{aligned}(\partial_x \sigma_{N+1})(0, h) &= (\partial_y R_N)(0, 0) + (\partial_x R_{N+1})(0, h) \\ &< (\partial_y R_N)(0, 0) + (\partial_x R_{N+1})(0, 0) \\ &= 0\end{aligned}$$

and

$$\begin{aligned}(\partial_x\sigma_{N+1})(h,h) &= (\partial_y R_N)(\tau_{N-1}(h),h) + (\partial_x R_{N+1})(h,h)\\ &> (\partial_y R_N)(h,h) + (\partial_x R_{N+1})(h,h)\\ &= 0.\end{aligned} \tag{2.23}$$

Hence, by continuity, there exists an $\eta \in (0,h)$ such that

$$(\partial_x\sigma_{N+1})(\eta,h) = 0. \tag{2.24}$$

For any such η, the first component of the right-hand side of (2.22) vanishes, the second component is nonnegative by (2.8), the third component is positive by (2.6) and the induction hypotheses (2.12) and (2.15), and the last component is positive by (2.6) and (2.21). Consequently, for such an η,

$$\left(\partial_x^2\sigma_{N+1}\right)(\eta,h) > -(\partial_x\partial_y\sigma_{N+1})(\eta,h) > 0. \tag{2.25}$$

It follows that there is precisely one $\eta \in (0,h)$ for which (2.24) holds. Moreover, this corresponds to the minimum of $\sigma_{N+1}(\eta,h)$ with respect to $\eta \in [0,h]$.

Let $\tau_N(h)$ denote the unique number $\eta \in (0,h)$ from the previous paragraph so that

$$(\partial_x\sigma_{N+1})(\tau_N(h),h) = 0 \tag{2.26}$$

and

$$s_{N+1}(h) = \sigma_{N+1}(\tau_N(h),h). \tag{2.27}$$

By the implicit function theorem applied to (2.26), $\tau_N \in C^1(0,H)$. Furthermore, differentiating (2.26),

$$\left(\partial_x^2\sigma_{N+1}\right)(\tau_N(h),h)\tau_N'(h) + (\partial_x\partial_y\sigma_{N+1})(\tau_N(h),h) = 0.$$

Recalling from the previous paragraph that (2.25) holds for $\eta = \tau_N(h)$, this yields (2.12) for $j = N$, and therewith, $\tau_N \in C([0,H])$.

Combining (2.19) and (2.26) gives (2.16) for $j = N$. Differentiating (2.27) with respect to h yields

$$s_{N+1}'(h) = (\partial_x\sigma_{N+1})(\tau_N(h),h)\,\tau_N'(h) + (\partial_y\sigma_{N+1})(\tau_N(h),h)$$

for $0 < h < H$. Subsequently, (2.20) and (2.26) give (2.17) for $n = N+1$.

This shows that if the induction hypothesis holds for any integer $N \geq 1$, it holds for $N+1$ too. Herewith, the proof of the theorem is complete. □

Further to Theorem 1, we can state the following.

THEOREM 2. *Let s_n denote the minimum value of S_n with respect to $t_0, t_1, \ldots, t_n$ satisfying constraint* (1.1).

(i) *Then s_n is a strictly decreasing function of $n \geq 1$.*

(ii) *If there exists an integer $p \geq 1$ such that $R_{j+p} = R_j$ for all $j \geq 1$, then $s_n - s_{n+p}$ is a strictly decreasing function of $n \geq 1$.*

(iii) *If $\{\partial_x R_j\}_{j=1}^{\infty}$ or $\{\partial_y R_j\}_{j=1}^{\infty}$ is equicontinuous in $\overline{\Omega}$, then*

$$s_n \to \int_0^H (-\partial_x R_1)(t,t)\,dt \qquad \text{as } n \to \infty. \tag{2.28}$$

Proof. We return to the preceding induction argument in N and reinstate h in the notation of s_n. Since $\tau_N(H) \in (0,H)$ represents the minimum of $\sigma_{N+1}(\eta,H)$ with respect to $\eta \in [0,H]$, together (2.4), (2.18), and (2.27) for $h = H$ imply that

$$\begin{aligned} s_{N+1}(H) &= \sigma_{N+1}(\tau_N(H),H) \\ &< \sigma_{N+1}(H,H) \\ &= s_N(H). \end{aligned}$$

This verifies part (i). The key to part (ii) is the assertion that the additional hypothesis implies that

$$\tau_{j-p}(\eta) < \tau_j(\eta) \quad \text{for all } 0 < \eta \leq H \tag{2.29}$$

and $j \geq p$. For $j = p$, this assertion is immediate from (2.11), (2.12), and (2.15). Suppose next that it is true for $j = N-1 \geq p$. Then using (2.19) and (2.6) with $j = N$, we have

$$\begin{aligned} (\partial_x \sigma_{N+1})(\eta,h) &< (\partial_y R_N)(\tau_{N-p-1}(\eta),\eta) + (\partial_x R_{N+1})(\eta,h) \\ &= (\partial_y R_{N-p})(\tau_{N-p-1}(\eta),\eta) + (\partial_x R_{N-p+1})(\eta,h) \end{aligned}$$

for any $0 < \eta < h \leq H$. Hence, by (2.16) for $j = N-p$, there holds

$$(\partial_x \sigma_{N+1})\left(\tau_{N-p}(h),h\right) < 0.$$

In light of (2.23) and the fact that $\tau_N(h)$ is the unique number $\eta \in (0,h)$ for which (2.24) holds, this implies (2.29) for $j = N$ and $\eta = h$. Thus, by induction, assertion (2.29) is true for all $j \geq p$. Subsequently, (2.6), (2.17), and (2.29) imply that

$$\begin{aligned} s_n'(h) - s_{n+p}'(h) &= (\partial_y R_n)\left(\tau_{n-1}(h),h\right) - (\partial_y R_n)\left(\tau_{n+p-1}(h),h\right) \\ &> 0 \end{aligned} \tag{2.30}$$

for all $0 < h < H$ and $n \geq 1$. Hence, first, using (2.27) for $N = n$ and $N = n+p$, second, using the fact that $\sigma_{n+1}(\tau_n(H),H)$ is the minimum value of $\sigma_{n+1}(\eta,H)$ with respect to $0 < \eta < H$, third, using (2.18) for $N = n$ and $N = n+p$, and fourth, using (2.30), we deduce that

$$\begin{aligned} s_{n+1}(H) - s_{n+p+1}(H) &= \sigma_{n+1}(\tau_n(H),H) - \sigma_{n+p+1}(\tau_{n+p}(H),H) \\ &< \sigma_{n+1}(\tau_{n+p}(H),H) - \sigma_{n+p+1}(\tau_{n+p}(H),H) \\ &= s_n(\tau_{n+p}(H)) - s_{n+p}(\tau_{n+p}(H)) \\ &< s_n(H) - s_{n+p}(H) \end{aligned}$$

for any $n \geq 1$. This verifies part (ii) of the theorem. To obtain the final part of the theorem in the case that $\{\partial_x R_j\}_{j=1}^{\infty}$ is equicontinuous in $\overline{\Omega}$, we note that by (2.2) and (2.4),

$$S_n(t_0,t_1,\ldots,t_n) = \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \left(-\partial_x R_j\right)(t,t_j)\, dt \tag{2.31}$$

for all $t_0,t_1,\ldots,t_n$ satisfying (1.1). Hence, by (2.6),

$$S_n(t_0,t_1,\ldots,t_n) > \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \left(-\partial_x R_j\right)(t,t)\, dt \tag{2.32}$$

for all such $t_0, t_1, \dots, t_n$. This inequality can be simplified by the observation that (2.4) implies

$$\partial_x R_j + \partial_y R_j = 0 \quad \text{on } \overline{\Omega} \setminus \Omega \tag{2.33}$$

for every $1 \le j \le n$. Hence, invoking (2.7), it can be established that

$$\partial_x R_j = \partial_x R_1 \quad \text{on } \overline{\Omega} \setminus \Omega \quad \text{for all } 1 \le j \le n. \tag{2.34}$$

Substituting (2.34) in (2.32) yields

$$s_n(H) > \int_0^H (-\partial_x R_1)(t,t)\, dt \quad \text{for all } n \ge 1. \tag{2.35}$$

Simultaneously, by the assumption of equicontinuity, given any $\varepsilon > 0$, there exists a $\delta > 0$ such that

$$|(\partial_x R_j)(t, t_j) - (\partial_x R_j)(t,t)| < \varepsilon \quad \text{for all } |t_j - t| < \delta,$$

$0 \le t \le t_j \le H$, and $j \ge 1$. Subsequently, choosing n so large that $H/n < \delta$, setting $t_j = jH/n$ for $0 \le j \le n$ in (2.31), and applying (2.34), we obtain

$$\begin{aligned} s_n(H) &\le S_n(0, H/n, 2H/n, \dots, H) \\ &\le \int_0^H (-\partial_x R_1)(t,t)\, dt + \varepsilon H \end{aligned} \tag{2.36}$$

for all such n. In view of the arbitrariness of ε, (2.35) and (2.36) provide the desired conclusion. To arrive at the same conclusion in the case that $\{\partial_y R_j\}_{j=1}^{\infty}$ is equicontinuous in $\overline{\Omega}$, we replace (2.31) by

$$S_n(t_0, t_1, \dots, t_n) = \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} (\partial_y R_j)(t_{j-1}, t)\, dt$$

and use the fact that given any $\varepsilon > 0$, there exists a $\delta > 0$ such that

$$|(\partial_y R_j)(t_{j-1}, t) - (\partial_y R_j)(t,t)| < \varepsilon \quad \text{for all } |t_{j-1} - t| < \delta,$$

$0 \le t_{j-1} \le t \le H$, and $j \ge 1$. Arguing as in the previous case, we obtain (2.28) with $\partial_y R_1$ instead of $-\partial_x R_1$. This is equivalent to (2.28) by (2.33). □

Recall that in the models considered so far, which are of type (1.18) and (1.19) with R given by either (1.20) or (1.21), all the R_j's are identical. Therefore, by part (ii) of the above theorem, $s_n - s_{n+1}$ is strictly decreasing in $n \ge 1$. This means that s_n is strictly convex in n. Consequently, the smallest value of n that satisfies the inequality $s_{n+1} \ge s_n - K$ gives an optimum of the cost function. Moreover, if this number n is such that the inequality holds with strictness, it is the unique optimum. If it is such that the inequality holds with equality, then $n+1$ is the one and only other optimum.

The next theorem and its corollary provide estimates of the difference between components t_j of the optimum. These estimates, in turn, should make the univariate search for the t_j's more efficient.

Hypothesis 4. There exist integers $1 \le k < k + p \le n$ such that

$$\prod_{j=0}^{p-1} (\partial_y R_{k+j})(x_j, y_j) \le \prod_{j=1}^{p} (-\partial_x R_{k+j})(x_j, y_j) \tag{2.37}$$

for all $0 \le x_0 < y_0 = x_1 < y_1 = x_2 < \cdots < y_{p-1} = x_p < y_p \le H$, with $y_p - x_p = y_0 - x_0$.

THEOREM 3. *Let* $t_0, t_1, \ldots, t_n$ *be the minimum of* S_n *under constraint* (1.1).

(a) *Suppose that Hypothesis* 4 *holds. Then*

$$t_{k+p} - t_{k+p-1} \le t_k - t_{k-1}. \tag{2.38}$$

Moreover, if inequality (2.37) *is strict, then so too is* (2.38).

(b) *Suppose that Hypothesis* 4 *holds with inequality* (2.37) *reversed. Then*

$$t_{k+p} - t_{k+p-1} \ge t_k - t_{k-1}. \tag{2.39}$$

Moreover, if reversed inequality (2.37) *is strict, then so too is* (2.39).

Proof. By (2.14) and (2.16),

$$0 = (\partial_y R_j)(t_{j-1}, t_j) + (\partial_x R_{j+1})(t_j, t_{j+1}) \quad \text{for all } 1 \le j \le n-1.$$

Hence,

$$(\partial_y R_k)(t_{k-1}, t_k) = P \times (-\partial_x R_{k+p})(t_{k+p-1}, t_{k+p}), \tag{2.40}$$

where

$$P = \prod_{j=1}^{p-1} \left(\frac{-\partial_x R_{k+j}}{\partial_y R_{k+j}} \right) (t_{k+j-1}, t_{k+j})$$

for any $1 \le k < k + p \le n$. The proof is by reductio ad absurdum, based on (2.40). Suppose that (2.38) is false. Then $t_{k+p-1} + t_k - t_{k-1} < t_{k+p}$. So, applying (2.6) to the last term in (2.40), we have

$$(\partial_y R_k)(t_{k-1}, t_k) > P \times (-\partial_x R_{k+p})(t_{k+p-1}, t_{k+p-1} + t_k - t_{k-1}). \tag{2.41}$$

This contradicts Hypothesis 4. By the same token, supposing that $t_{k+p-1} + t_k - t_{k-1} \le t_{k+p}$, we obtain (2.41) with a weak inequality. This again contradicts Hypothesis 4, if (2.37) is strict. In contrast, if (2.39) is false, then $t_{k-1} < t_k - t_{k+p} + t_{k+p-1}$. Hence, applying (2.6) to the left-hand side of (2.40), we deduce that

$$(\partial_y R_k)(t_k - t_{k+p} + t_{k+p-1}, t_k) < P \times (-\partial_x R_{k+p})(t_{k+p-1}, t_{k+p}). \tag{2.42}$$

This contradicts Hypothesis 4 if inequality (2.37) is reversed. Similarly, if $t_{k-1} \le t_k - t_{k+p} + t_{k+p-1}$, then (2.42) holds with a weak inequality, which contradicts Hypothesis 4 if inequality (2.37) is reversed and strict. To summarize, in all cases, by supposing the negation of the inequality that we wish to prove, we arrive at a contradiction of the inequality of type (2.37) that we have presupposed. □

COROLLARY 3.1. *Suppose that* $n = mp + r$ *for some integers* $m \ge 1$, $p \ge 1$, *and* $0 \le r \le p - 1$.

(a) *When* (2.38) *holds for every* $1 \le k \le n-p$, *then*

$$t_{j+p} - t_j \le \frac{H}{m} \tag{2.43}$$

for $(m-1)p \le j \le n-p$, *with strict inequality if* $r \ge 1$ *or* (2.38) *is strict for some* $1 \le k \le n-p$.

(b) *When* (2.39) *holds for every* $1 \le k \le n-p$, *then* (2.43) *holds for* $0 \le j \le r$, *with strict inequality if* $r \ge 1$ *or* (2.39) *is strict for some* $1 \le k \le n-p$.

Proof. Summing (2.38) from $k = j + (i-m)p + 1$ to $k = j + (i+1-m)p$ gives

$$t_{j+(i+2-m)p} - t_{j+(i+1-m)p} \le t_{j+(i+1-m)p} - t_{j+(i-m)p}$$

for any $1 \le i \le m-1$ and $(m-1)p \le j \le n-p$, with strict inequality if (2.38) is strict for some k. Consequently,

$$t_{j+p} - t_j \le \frac{1}{m}\sum_{i=1}^{m}\left(t_{j+(i+1-m)p} - t_{j+(i-m)p}\right) = \frac{t_{j+p} - t_{j-(m-1)p}}{m},$$

with strict inequality if (2.38) is strict for some $j-(m-1)p+1 \le k \le j$. This yields (2.43). Moreover, since $t_{j+p} < H$ for $j+p \le n-1$, and $t_{j-(m-1)p} > 0$ for $j-(m-1)p \ge 1$, it yields (2.43) with strictness if $j \le n-p-1$ or $j \ge (m-1)p+1$. On the other hand, the only situation when the latter is not the case is when $j = n-p = (m-1)p$, in which situation, by what has been noted previously, there is strictness in (2.43) if (2.38) is strict for some $1 \le k \le n-p$. This gives part (a). The proof of part (b) is analogous. □

3. Specific models.

The goal of this section is to check that the models in hand conform to the general picture in the previous section. The specific aim is, thus, to verify that respective functions (1.20) and (1.21) satisfy generic Hypothesis 1 and to identify their prospects for satisfying Hypothesis 2. Theorem 1 will then yield the optimality result for fixed n. The convexity of the optimal cost function with respect to n follows immediately hereafter, via Theorem 2 with $p = 1$ in part (ii).

3.1. Preliminaries.

To realize the intended goal, let us first establish some additional properties of the function κ defined by (1.11). For a start, dropping the argument (x, y) from derivatives, it can be calculated that

$$\partial_x \kappa = -D(x)F(\gamma(x,y)) - \theta(x)\kappa(x,y) \tag{3.1}$$

and

$$\partial_y \kappa = D(y)\rho(x,y)F(\gamma(x,y)), \tag{3.2}$$

where

$$\gamma(x,y) = e^{\Theta(x)}\kappa(x,y) \quad \text{and} \quad \rho(x,y) = e^{\Theta(y)-\Theta(x)}.$$

For further convenience, we set

$$\Phi(t) = \frac{D'(t)}{D(t)} + \theta(t), \tag{3.3}$$

$$q(x,y) = e^{\Theta(y)}D(y) - e^{\Theta(x)}D(x), \tag{3.4}$$

and

$$\Psi(v) = \int_0^v \frac{dz}{F(z)} \tag{3.5}$$

for $0 \le v \le \ell$. Noting that Ψ is the inverse of ψ, (1.11) may be reformulated as

$$\Psi(\gamma(x, y)) = \mu(x, y). \tag{3.6}$$

With this notation, we have the following.

Lemma 1. *If Φ is nonincreasing on $(0, H)$, then*

$$\Phi(x)\Psi(\gamma(x, y)) \ge q(x, y) \ge \Phi(y)\Psi(\gamma(x, y)) \tag{3.7}$$

for all $0 < x < y < H$.

Proof. There holds

$$\begin{aligned} q(x, y) &= \int_x^y \partial_t \left\{ e^{\Theta(t)} D(t) \right\} \, dt \\ &= \int_x^y \left\{ \theta(t) \, e^{\Theta(t)} D(t) + e^{\Theta(t)} D'(t) \right\} \, dt \\ &= \int_x^y \Phi(t) \, e^{\Theta(t)} D(t) \, dt \end{aligned}$$

for all $0 < x < y < H$. So, if Φ is nonincreasing,

$$q(x, y) \ge \int_x^y \Phi(y) \, e^{\Theta(t)} D(t) \, dt = \Phi(y)\mu(x, y).$$

This yields the right-hand inequality in (3.7) via (3.6). The proof of the left-hand inequality is analogous. □

Lemma 2. *If $D = e^{-\Theta}$, then $\Phi \equiv 0$.*

Lemma 3. *If $F(v) = v^\beta$ for some $\beta < 1$, then*

$$\frac{\Psi(v)F(v)}{v} = 1 + \Psi(v)F'(v) = \frac{1}{1 - \beta} \qquad \textit{for } v > 0.$$

The proofs of the last two lemmata are left as an exercise.

3.2. The OHD cost function. In the OHD model, R is given by (1.20), where w is given by (1.15). As it turns out, the specific form of w is not so important. It suffices to assume that $w \in C([0, H]) \cap C^1(0, H)$ is positive on $(0, H)$. We establish the following.

Lemma 4. *Let R be given by* (1.20) *with κ given by* (1.11)*. Suppose that $F \in C^1(0, \ell)$ or $D = e^{-\Theta}$. Then R satisfies Hypothesis* 1*, and $\partial_x R = \partial_y R = 0$ on $\overline{\Omega} \setminus \Omega$.*

Proof. Direct calculation reveals that

$$\partial_x R = -w(x)\kappa(x, y) \tag{3.8}$$

and

$$\partial_y R = D(y)Q(x, y), \tag{3.9}$$

where

$$Q(x,y) = \int_x^y w(t)\rho(t,y)F(\gamma(t,y))\,dt. \tag{3.10}$$

The differentiation under the integral sign implicit in the computation of (3.9) can be justified by changing the variable of integration in (1.20) to $\int_t^y e^{\Theta(u)}D(u)\,du$. Further calculation gives

$$\partial_x^2 R = -w'(x)\kappa(x,y) + w(x)\{D(x)F(\gamma(x,y)) + \theta(x)\kappa(x,y)\}, \tag{3.11}$$

$$\partial_x\partial_y R = -w(x)D(y)\rho(x,y)F(\gamma(x,y)), \tag{3.12}$$

and

$$\partial_y^2 R = D'(y)Q(x,y) + D(y)\left(\partial_y Q\right)(x,y).$$

To evaluate the last expression, we rewrite

$$Q(x,y) = \int_0^{y-x} w(y-u)\rho(y-u,y)F(\gamma(y-u,y))\,du.$$

When $F \in C^1(0,\ell)$, this formally yields

$$\begin{aligned}
\partial_y Q = {} & w(x)\rho(x,y)F(\gamma(x,y)) \\
& + \int_0^{y-x} w'(y-u)\rho(y-u,y)F(\gamma(y-u,y))\,du \\
& - \int_0^{y-x} w(y-u)\rho(y-u,y)\theta(y-u)F(\gamma(y-u,y))\,du \\
& + \int_0^{y-x} w(y-u)\rho(y-u,y)\theta(y)F(\gamma(y-u,y))\,du \\
& + \int_0^{y-x} w(y-u)\rho(y-u,y)\left(FF'\right)(\gamma(y-u,y))q(y-u,y)\,du \\
= {} & w(x)\rho(x,y)F(\gamma(x,y)) \\
& + \int_x^y [w'(t) + w(t)\{\theta(y)-\theta(t)\}]\rho(t,y)F(\gamma(t,y))\,dt \\
& + \int_x^y w(t)\rho(t,y)q(t,y)\left(FF'\right)(\gamma(t,y))\,dt.
\end{aligned} \tag{3.13}$$

Rigorous justification follows once again by a change of variable similar to that used for (3.9). As in that case, since otherwise this only complicates the calculations, we omit the details. When $D = e^{-\Theta}$, it can be determined that $\gamma(y-u,y) = \psi(u)$. So the terms in (3.13) involving F' do not appear. The lemma is an immediate consequence of the above analysis. □

LEMMA 5. *Suppose, in addition to the hypotheses of Lemma* 4, *that* Φ *is nonincreasing on* $(0,H)$. *Then if* $D = e^{-\Theta}$ *or* $F(v) = v^\beta$ *for some* $\beta < 1$, *there holds* $\mathcal{L}_x R \geq 0$ *in* Ω, *where*

$$f(t) = \theta(t) - \frac{w'(t)}{w(t)}$$

or

$$f(t) = \theta(t) - \frac{w'(t)}{w(t)} - \frac{\Phi(t)}{1-\beta},$$

respectively.

Proof. Embroidering on the proof of Lemma 4, using (3.8), (3.11), and (3.12), and tidying up, it can be deduced that for any operator of the form (2.9), there holds

$$\mathcal{L}_x R = w(x)\kappa(x,y)\phi(x,y),$$

where

$$\phi(x,y) = \theta(x) - \frac{w'(x)}{w(x)} - q(x,y)\frac{F(\gamma(x,y))}{\gamma(x,y)} - f(x).$$

If Φ is nonincreasing, Lemma 1 implies that

$$\phi(x,y) \geq \theta(x) - \frac{w'(x)}{w(x)} - \Phi(x)\frac{\Psi(v)F(v)}{v} - f(x),$$

where $v = \gamma(x,y)$. In light of Lemmata 2 and 3, this gives the asserted. □

LEMMA 6. *Suppose, in addition to the hypotheses of Lemma* 5, *that $w'/w - \theta$ is nonincreasing on* $(0,H)$. *Then the conclusions of Lemma* 5 *apply to $\mathcal{L}_y R$ as well as $\mathcal{L}_x R$.*

Proof. When $F \in C^1(0,\ell)$, combining (3.9) and (3.12)–(3.13) and eliminating D' using (3.3), it can be calculated that for any operator of the form (2.10), there holds

$$\mathcal{L}_y R = D(y)\int_x^y w(t)\rho(t,y)F(\gamma(t,y))\phi(t,y)\,dt,$$

where

$$\phi(x,y) = f(y) + \frac{w'(x)}{w(x)} - \theta(x) + \Phi(y) + q(x,y)F'(\gamma(x,y)). \tag{3.14}$$

Now, if $w'/w - \theta$ and Φ are nonincreasing,

$$\phi(x,y) \geq f(y) + \frac{w'(y)}{w(y)} - \theta(y) + \Phi(y)\left\{1 + \Psi(v)F'(v)\right\},$$

where $v = \gamma(x,y)$. Recalling Lemmata 2 and 3, the assertion follows. When $D = e^{-\Theta}$, the proof is simpler, since $\Phi \equiv 0$ by Lemma 2, and the term involving F' does not appear in (3.14). □

LEMMA 7. *Let R be given by* (1.20) *and* $(x,y) \in \Omega$.

(i) *Suppose that $D = e^{-\Theta}$. Then*

$$\left(-\partial_x R\right)(x,y) = w(x)D(x)\psi(y-x). \tag{3.15}$$

Furthermore, if wD is nondecreasing on $(0,H)$, *then*

$$\left(\partial_y R\right)(x,y) \leq w(y)D(y)\psi(y-x), \tag{3.16}$$

with equality if and only if wD is constant on (x,y). *On the other hand, if wD is nonincreasing on* $(0,H)$, *then* (3.16) *holds with the inequality sign reversed, and equality as stated before.*

(ii) *Suppose that* $F(v) = v^\beta$ *for some* $0 \leq \beta < 1$. *Set*

$$A(t) = w(t)\left\{D(t)\, e^{\beta\Theta(t)}\right\}^{1/(1-\beta)}. \tag{3.17}$$

If there exists a number λ *such that* $De^{\lambda\Theta}$, $we^{\{(\beta-\lambda)/(1-\beta)\}\Theta}$, *and* $(1-\lambda)\theta$ *are nondecreasing on* $(0, H)$, *then*

$$\left(-\partial_x R\right)(x,y) \geq A(x)\psi\left(\textstyle\int_0^{y-x} e^{(1-\lambda)\theta(x)u}\, du\right) \tag{3.18}$$

and

$$\left(\partial_y R\right)(x,y) \leq A(y)\psi\left(\textstyle\int_0^{y-x} e^{(1-\lambda)\theta(y)u}\, du\right), \tag{3.19}$$

with equality in both simultaneously if and only if $A^{1-\beta} + (1-\lambda)\theta$ *is constant on* (x, y). *On the other hand, if there exists a number* λ *such that the aforesaid functions are nonincreasing on* $(0, H)$, *then* (3.18) *and* (3.19) *hold with the inequality sign reversed, and equality in both simultaneously as stated before.*

Proof. By (1.11) and (3.8),

$$\left(-\partial_x R\right)(x,y) = w(x)\, e^{-\Theta(x)}\psi(\mu(x,y)), \tag{3.20}$$

while by (1.9), (3.6), (3.9), and (3.10),

$$\left(\partial_y R\right)(x,y) = D(y)\int_x^y w(t)\rho(t,y)\psi'(\mu(t,y))\, dt. \tag{3.21}$$

Now, if $D = e^{-\Theta}$, one has $\mu(x,y) = y - x$, and (3.20) reduces to (3.15), while (3.21) reduces to

$$\left(\partial_y R\right)(x,y) = \int_x^y w(t)D(t)\psi'(y-t)\, dt.$$

Hence, if wD is nondecreasing,

$$\left(\partial_y R\right)(x,y) \leq w(y)D(y)\int_x^y \psi'(y-t)\, dt.$$

This gives (3.16). Moreover, since $wD \in C([0, H])$, it gives (3.16) with equality if and only if wD is constant on (x, y). Thus, the first and second assertions of part (i) are proven. The outstanding assertion is proven similarly to the second. For part (ii), we note that when $F(v) = v^\beta$ for some $0 \leq \beta < 1$, the function ψ is given by (1.12). Hence,

$$\psi'(z) = \{(1-\beta)z\}^{\beta/(1-\beta)} \tag{3.22}$$

for $z > 0$. Furthermore, if $De^{\lambda\Theta}$ is nondecreasing on $(0, H)$ for some λ, then

$$\begin{aligned} D(t)\, e^{\Theta(t)}\int_t^y \rho^{1-\lambda}(t,u)\, du &\leq \mu(t,y) \\ &\leq D(y)\, e^{\Theta(y)}\rho^{\lambda-1}(t,y)\int_t^y \rho^{1-\lambda}(t,u)\, du \end{aligned} \tag{3.23}$$

for all $x \le t < y$. Thus,

$$(-\partial_x R)(x,y) \ge A(x)\psi\left(\textstyle\int_x^y \rho^{1-\lambda}(x,t)\,dt\right), \tag{3.24}$$

with equality if and only if $De^{\lambda\Theta}$ is constant on (x,y), while if $W = we^{\{(\beta-\lambda)/(1-\beta)\}\Theta}$ is nondecreasing too,

$$(\partial_y R)(x,y) \le A(y)\int_x^y \rho^{1-\lambda}(t,y)\psi'\left(\textstyle\int_t^y \rho^{1-\lambda}(t,u)\,du\right)\,dt, \tag{3.25}$$

with equality if $De^{\lambda\Theta}$ and W are constant on (x,y) and only if W is constant on (x,y). To proceed, we use the observation that if $(1-\lambda)\theta$ is nondecreasing, then

$$\rho^{1-\lambda}(x,t) \ge e^{(1-\lambda)\theta(x)(t-x)} \quad \text{for all } x < t \le y, \tag{3.26}$$

while

$$\rho^{1-\lambda}(t,u) \le e^{(1-\lambda)\theta(y)(u-t)} \quad \text{for all } x \le t < u \le y. \tag{3.27}$$

Substituting (3.26) in (3.24) and simplifying yields (3.18) with equality if and only if $De^{\lambda\Theta}$ and $(1-\lambda)\theta$ are constant on $(0,H)$. Whereas, using (3.27) to eliminate both occurrences of $\rho^{1-\lambda}$ from (3.25) and simplifying the resulting inequality yields (3.19) with equality if $De^{\lambda\Theta}$, W and $(1-\lambda)\theta$ are constant on (x,y), and only if W and $(1-\lambda)\theta$ are constant on (x,y). Noting that $A^{1-\beta} = De^{\lambda\Theta}W^{1-\beta}$, this completes the proof of the first assertion of part (ii). The proof of the second assertion is entirely analogous. □

Theorems 1 and 2 and the first three of the above lemmata yield the following central result for the OHD model.

THEOREM 4. *Consider the model* (1.18)–(1.20), *under the introductory assumptions. Suppose furthermore that $D'/D+\theta$ and $w'/w-\theta$ are nonincreasing on $(0,H)$. Then, for every $n \ge 1$, S_n has a unique minimum with respect to $t_0,\dots,t_n$ satisfying* (1.1) *when $D = e^{-\Theta}$ or $F(v) = v^\beta$ for some $\beta < 1$. Moreover, if s_n denotes the minimum value of S_n under constraint* (1.1), *then $n \mapsto s_n$ is strictly decreasing and strictly convex.*

Theorem 4 immediately says the following about the optimum of the total cost.

COROLLARY 4.1. *The following alternatives are mutually exclusive.*

(i) *If $K > s_1 - s_2$, then $\mathcal{C}$ has a unique minimum for $n = 1$.*

(ii) *If there exists a number $N \ge 2$ such that $s_{N-1} - s_N > K > s_N - s_{N+1}$, then $\mathcal{C}$ has a unique minimum for $n = N$.*

(iii) *If there exists a number $N \ge 1$ such that $K = s_N - s_{N+1}$, then $\mathcal{C}$ has precisely two minima: for $n = N$ and $n = N+1$.*

Since the numbers s_n for $n \ge 1$ are independent of K, each of the possibilities in the above corollary is viable.

Theorem 3 and Lemma 7 supplement Theorem 4 with the following.

THEOREM 5. *Let $t_0, t_1, \dots, t_n$ be the minimum of S_n under constraint* (1.1) *given by Theorem 4 for some $n \ge 2$.*

(i) *Suppose that $D = e^{-\Theta}$. If wD is nondecreasing on $(0,H)$, then*

$$t_{j+1} - t_j \le t_j - t_{j-1} \quad \textit{for all } 1 \le j \le n-1. \tag{3.28}$$

Moreover, the inequality is strict if wD is strictly increasing. On the other hand, if wD is nonincreasing on $(0,H)$, then

$$t_{j+1} - t_j \ge t_j - t_{j-1} \quad \textit{for all } 1 \le j \le n-1. \tag{3.29}$$

Moreover, the inequality is strict if wD is strictly decreasing.

(ii) *Suppose that* $F(v) = v^\beta$ *for some* $0 \le \beta < 1$. *If there exists a number* λ *such that* $De^{\lambda\Theta}$, $we^{\{(\beta-\lambda)/(1-\beta)\}\Theta}$, *and* $(1-\lambda)\theta$ *are nondecreasing on* $(0,H)$, *then* (3.28) *holds. Moreover, the inequality is strict if* $Dw^{1-\beta}e^{\beta\theta} + (1-\lambda)\theta$ *is strictly increasing. On the other hand, if there exists a number* λ *such that the aforesaid functions are nonincreasing on* $(0,H)$, *then* (3.29) *holds. Moreover, the inequality is strict if* $Dw^{1-\beta}e^{\beta\theta} + (1-\lambda)\theta$ *is strictly decreasing.*

Corollary 5.1. *If* $D = e^{-\Theta}$ *and* wD *is constant on* $(0,H)$, *or if* $F(v) = v^\beta$ *for some* $0 \le \beta < 1$, *and there exists a number* λ *such that* $De^{\lambda\Theta}$, $w\,e^{\{(\beta-\lambda)/(1-\beta)\}\Theta}$ *and* $(1-\lambda)\theta$, *are constant on* $(0,H)$, *then*

$$t_{j+1} - t_j = t_j - t_{j-1} \quad \textit{for all } 1 \le j \le n-1. \tag{3.30}$$

Note that, in itself, the assumption that θ is nonnegative is not needed in Theorem 4, its corollary, nor for Theorem 5. The essential assumption is that w is positive on $(0,H)$.

Recalling the equivalence of the OHD and OHP cost functions when the inventory dynamics are described by classical linear model (1.2), Theorem 4 improves on previous results on the existence of an optimal replenishment schedule for this model [2, 5, 9, 12] for both cost functions. It disposes of the diverse assumptions imposed on the monotonicity of the demand rate D and reduces the various assumptions on the deterioration rate θ to the single supposition that $w'/w - \theta$ is nonincreasing. Regarding nonlinear model (1.3), supposing that $\beta < 1$, G satisfies the same introductory hypotheses as D and the basic hypotheses on θ are unaltered, Theorem 4 establishes the existence of an optimal replenishment schedule for the OHD cost function when $G'/G + (1-\beta)\theta$ and $w'/w - \theta$ are nonincreasing on $(0,H)$. Superficially, it would appear that there have been no comparable previous results. Nevertheless, some can be distilled from those obtained in [2] for the OHP cost function, because when θ is constant, the cost of deterioration is a constant multiple $c_2\theta/c_1$ of the cost of holding, while the analysis in [2] carries through even if $c_2 = 0$. Theorem 4 improves on these inferred results by relaxing the presupposition that $0 \le \beta < 1$ to $\beta < 1$, by disposing of the assumption that G' is positive, and by removing the assumption that θ is constant, besides shedding a few other inessential technical hypotheses. Theorem 5 correspondingly generalizes and sharpens prior results on the successive decrement or increment of the length of a replenishment period for linear model (1.2) and nonlinear model (1.3). Over and above this, Theorems 4 and 5 cover a class of models not previously considered, viz.,

$$I'(t) = -e^{-\Theta(t)}F\left(e^{\Theta(t)}I(t)\right) - \theta(t)I(t), \tag{3.31}$$

in which $F \in C(0,\ell)$, $F(v) > 0$ for $0 < v < \ell$, and (1.6) with $\mu(0,H) = H$ holds for some $\ell > 0$, of which (1.4) is a special case. For this class of models, an optimal replenishment schedule has been shown to exist under the single supposition that $w'/w - \theta$ is nonincreasing. Successive replenishment periods within this schedule have been shown to be shorter or longer dependent upon an auxiliary condition concerning the monotonicity of $we^{-\Theta}$.

3.3. The OHP cost function. The OHP model is (1.18), (1.19), and (1.21) where κ is given by (1.11) and c_1 and c_2 are positive constants. Since the first term on the right-hand side of (1.21) may be viewed as a special case of (1.20) (with $w \equiv c_1$), to determine when Theorem 1 can be applied to (1.21), it suffices to examine the influence of the second term.

We prove the following.

LEMMA 8. *Let κ be given by* (1.11). *Suppose that $F \in C([0,\ell]) \cap C^1(0,\ell)$ is nondecreasing. Then κ satisfies Hypothesis* 1 *with weak inequality in* (2.6)*, and $\partial_y \kappa = -\partial_x \kappa = D(x)F(0)$ on $\overline{\Omega} \setminus \Omega$.*

Proof. Differentiating (3.1) and (3.2), one can deduce that

$$\begin{aligned} \partial_x^2 \kappa &= \{D(x)\theta(x) - D'(x)\}F(\gamma(x,y)) \\ &\quad + D^2(x)\, e^{\Theta(x)}\, (FF')\,(\gamma(x,y)) + \left\{\theta^2(x) - \theta'(x)\right\} \kappa(x,y), \end{aligned} \tag{3.32}$$

$$\begin{aligned} \partial_x \partial_y \kappa &= -D(y)\theta(x)\rho(x,y)F(\gamma(x,y)) \\ &\quad - D(x)D(y)\, e^{\Theta(y)}\, (FF')\,(\gamma(x,y)), \end{aligned} \tag{3.33}$$

and

$$\begin{aligned} \partial_y^2 \kappa &= \{D'(y) + D(y)\theta(y)\}\rho(x,y)F(\gamma(x,y)) \\ &\quad + D^2(y)\, e^{\Theta(y)} \rho(x,y)\, (FF')\,(\gamma(x,y)). \end{aligned} \tag{3.34}$$

Together with (3.1) and (3.2), these relations yield the lemma. □

LEMMA 9. *Suppose, in addition to the hypotheses of Lemma* 8*, that Φ and θ are nonincreasing on $(0,H)$. Then if $D = e^{-\Theta}$ or $F(v) = v^\beta$ for some $0 \le \beta < 1$, there holds $\mathcal{L}_x \kappa \ge 0$ in Ω, where*

$$f(t) = \theta(t) \tag{3.35}$$

or

$$f(t) = \theta(t) - \frac{\Phi(t)}{1-\beta} \tag{3.36}$$

respectively.

Proof. Substituting (3.1), (3.32), and (3.33) in (2.9) and using (3.3) to eliminate D', one finds

$$\mathcal{L}_x \kappa = D(x)F(\gamma(x,y))\phi_1(x,y) + \kappa(x,y)\phi_2(x,y)$$

for any function f, where

$$\phi_1(x,y) = \theta(x) - \Phi(x) - q(x,y)F'(\gamma(x,y)) - f(x)$$

and

$$\phi_2(x,y) = \theta^2(x) - \theta'(x) - \theta(x)q(x,y)\frac{F(\gamma(x,y))}{\gamma(x,y)} - \theta(x)f(x).$$

By Lemma 1, if Φ is nonincreasing on $(0,H)$, then

$$\phi_1(x,y) \ge \theta(x) - \Phi(x)\left\{1 + \Psi(v)F'(v)\right\} - f(x),$$

with $v = \gamma(x,y)$. Similarly,

$$\phi_2(x,y) \ge \theta^2(x) - \theta'(x) - \theta(x)\Phi(x)\frac{\Psi(v)F(v)}{v} - \theta(x)f(x). \tag{3.37}$$

Hence, by Lemmata 2 and 3, if, furthermore, $D = e^{-\Theta}$ and f is given by (3.35), or $F(v) = v^\beta$ for some $0 \le \beta < 1$, and f is given by (3.36), one has $\phi_1 \ge 0$ in Ω.

Moreover, in either of these cases, (3.37) reduces to $\phi_2(x,y) \geq -\theta'(x)$. Thus, $\mathcal{L}_x\kappa \geq 0$ when θ is also nonincreasing on $(0,H)$. □

LEMMA 10. *Suppose, in addition to the hypotheses of Lemma* 8, *that* Φ *is nonincreasing and* θ *is nondecreasing on* $(0,H)$. *Then, the conclusions of Lemma* 9 *apply with* $\mathcal{L}_y\kappa$ *instead of* $\mathcal{L}_x\kappa$.

Proof. Substituting (3.2), (3.33), and (3.34) in (2.10), one obtains

$$\mathcal{L}_y\kappa = D(y)\rho(x,y)F(\gamma(x,y))\phi(x,y),$$

in which

$$\phi(x,y) = f(y) - \theta(x) + \Phi(y) + q(x,y)F'(\gamma(x,y)).$$

By Lemma 1, when Φ is nonincreasing and θ is nondecreasing on $(0,H)$,

$$\phi(x,y) \geq f(y) - \theta(y) + \Phi(y)\left\{1 + \Psi(v)F'(v)\right\},$$

with $v = \gamma(x,y)$. This gives $\mathcal{L}_y\kappa \geq 0$ under the conditions stated, via Lemmata 2 and 3. □

LEMMA 11. *Let the assumptions of Lemma* 8 *hold, and* $(x,y) \in \Omega$.

(i) *Suppose that* $D = e^{-\Theta}$. *Then*

$$(-\partial_x\kappa)(x,y) = D(x)\{\psi'(y-x) + \theta(x)\psi(y-x)\}. \tag{3.38}$$

Furthermore, if θ *is nonincreasing on* $(0,H)$, *then*

$$(\partial_y\kappa)(x,y) \geq D(y)\{\psi'(y-x) + \theta(y)\psi(y-x)\}. \tag{3.39}$$

(ii) *Suppose that* $F(v) = v^\beta$ *for some* $0 \leq \beta < 1$. *Let* $A(t)$ *be given by* (3.17) *with* $w(t) = 1$. *If* $De^{\beta\Theta}$ *and* θ *are nondecreasing on* $(0,H)$, *then*

$$(-\partial_x\kappa)(x,y) \geq A(x)\, e^{(1-\beta)\theta(x)(y-x)}\psi'\left(\textstyle\int_0^{y-x} e^{(1-\beta)\theta(x)u}\, du\right) \tag{3.40}$$

and

$$(\partial_y\kappa)(x,y) \leq A(y)\, e^{(1-\beta)\theta(y)(y-x)}\psi'\left(\textstyle\int_0^{y-x} e^{(1-\beta)\theta(y)u}\, du\right). \tag{3.41}$$

On the other hand, if $De^{\beta\Theta}$ *and* θ *are nonincreasing on* $(0,H)$, *then* (3.40) *and* (3.41) *hold with the inequality sign reversed.*

Proof. Irrespective of any extraordinary assumptions, by (1.9), (1.11), (3.1), and (3.6),

$$(-\partial_x\kappa)(x,y) = D(x)\psi'(\mu(x,y)) + \theta(x)\, e^{-\Theta(x)}\psi(\mu(x,y)), \tag{3.42}$$

while by (1.9), (3.2), and (3.6),

$$(\partial_y\kappa)(x,y) = D(y)\rho(x,y)\psi'(\mu(x,y)). \tag{3.43}$$

Consequently, if $D = e^{-\Theta}$, (3.42) reduces to (3.38). This gives the first assertion of part (i). To prove the remaining assertion of part (i), we note that when θ is nonincreasing, (3.43) yields

$$(\partial_y\kappa)(x,y) \geq D(y)\, e^{\theta(y)(y-x)}\psi'(y-x).$$

Consequently, to prove (3.39), it suffices to show that

$$e^{\theta(y)(y-x)}\psi'(y-x) \geq \psi'(y-x) + \theta(y)\psi(y-x).$$

Setting $v = \psi(y-x)$ and using (1.9) and (3.5), the above inequality is equivalent to

$$e^{\theta(y)\Psi(v)} \geq 1 + \theta(y)v/F(v). \tag{3.44}$$

However, since $e^{\eta} \geq 1+\eta$ for any number η, there holds $e^{\theta(y)\Psi(v)} \geq 1 + \theta(y)\Psi(v)$ for all $0 < v \leq \ell$. In turn, if F is nondecreasing on $[0,\ell]$, by (3.5) we have $\Psi(v) \geq v/F(v)$ for any such v. Putting these two inequalities together confirms (3.44) under the assumptions of Lemma 8, and therewith, (3.39). To prove part (ii), we note that if $De^{\beta\Theta}$ is nondecreasing on $(0,H)$, then (3.23) with $\lambda = \beta$ holds for all $x \leq t < y$. Hence, recalling (1.12), (3.22), (3.42), and (3.43),

$$\left(-\partial_x\kappa\right)(x,y) \geq A(x)\left\{\psi'\left(\int_x^y \rho^{1-\beta}(x,t)\,dt\right) + \theta(x)\psi\left(\int_x^y \rho^{1-\beta}(x,t)\,dt\right)\right\}$$

and

$$\left(\partial_y\kappa\right)(x,y) \leq A(y)\rho^{1-\beta}(x,y)\psi'\left(\int_x^y \rho^{1-\beta}(x,t)\,dt\right).$$

In addition, if θ is nondecreasing too, then (3.26) and (3.27) with $\lambda = \beta$ hold too. Combining these inequalities yields (3.40) and (3.41). This confirms the first assertion of part (ii). The proof of the remaining assertion mimics that of the first. □

From Theorems 1–3 and the four lemmata above, we obtain the following.

THEOREM 6. *Consider the model* (1.18), (1.19), *and* (1.21), *under the introductory assumptions. Suppose furthermore that D'/D is nonincreasing on $(0,H)$, θ is constant, and $F \in C([0,\ell]) \cap C^1(0,\ell)$ is nondecreasing. Then for every $n \geq 1$, S_n has a unique minimum with respect to $t_0,\ldots,t_n$ satisfying* (1.1) *when $D = e^{-\Theta}$ or $F(v) = v^{\beta}$ for some $0 \leq \beta < 1$. Moreover, if s_n denotes the minimum value of S_n under constraint* (1.1), *then $n \mapsto s_n$ is strictly decreasing and strictly convex.*

COROLLARY 6.1. *Verbatim, the conclusions of Corollary* 4.1 *hold.*

THEOREM 7. *Let $t_0, t_1, \ldots, t_n$ be the minimum of S_n under constraint* (1.1) *given by Theorem* 6 *for some $n \geq 2$.*

(i) *Suppose that $D = e^{-\Theta}$. Then* (3.29) *holds. Moreover, the inequality is strict if and only if $\theta > 0$.*

(ii) *Suppose that $F(v) = v^{\beta}$ for some $0 \leq \beta < 1$. If $De^{\beta\Theta}$ is nondecreasing on $(0,H)$, then* (3.28) *holds. Moreover, the inequality is strict if $De^{\beta\Theta}$ is strictly increasing. On the other hand, if $De^{\beta\Theta}$ is nonincreasing on $(0,H)$, then* (3.29) *holds. Moreover, the inequality is strict if $De^{\beta\Theta}$ is strictly decreasing.*

COROLLARY 7.1. *If $D = e^{-\Theta}$ and $\theta = 0$, or if $F(v) = v^{\beta}$ for some $0 \leq \beta < 1$, and $De^{\beta\Theta}$ is constant on $(0,H)$, then* (3.30) *holds.*

For model (1.3) with $0 \leq \beta < 1$ and θ a nonnegative constant, Theorem 6 recovers the existence result found in [2]. For, in this case, $F(v) = v^{\beta}$, and the supposition that D'/D is nonincreasing is equivalent to the hypothesis that G'/G is nonincreasing. In fact, the theorem sharpens the previous result by discarding the explicit assumption that $G' > 0$ and removing some inessential implicit assumptions in [2]. Theorem 7 generalizes the conclusion that if $G' > 0$, then successive replenishment periods within the optimal schedule are strictly decreasing in a number of ways. Apart from this, Theorems 6 and 7 provide results on model (3.31), where $F \in C([0,\ell]) \cap C^1(0,\ell)$, $F(v) > 0$ and $F'(v) \geq 0$ for $0 < v < \ell$, (1.6), with $\mu(0,H) = H$ holds for some $\ell > 0$, and θ is constant. Such a model has not been considered before.

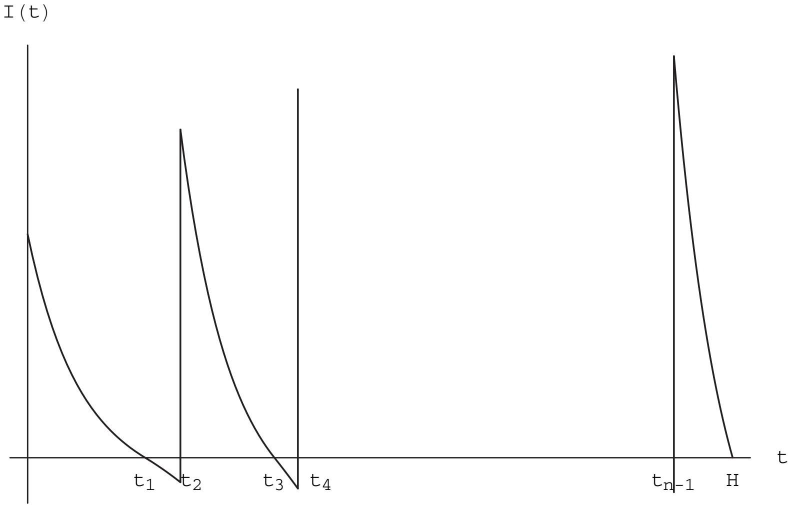


FIG. 2. *Typical inventory behavior with shortage for model CC.*

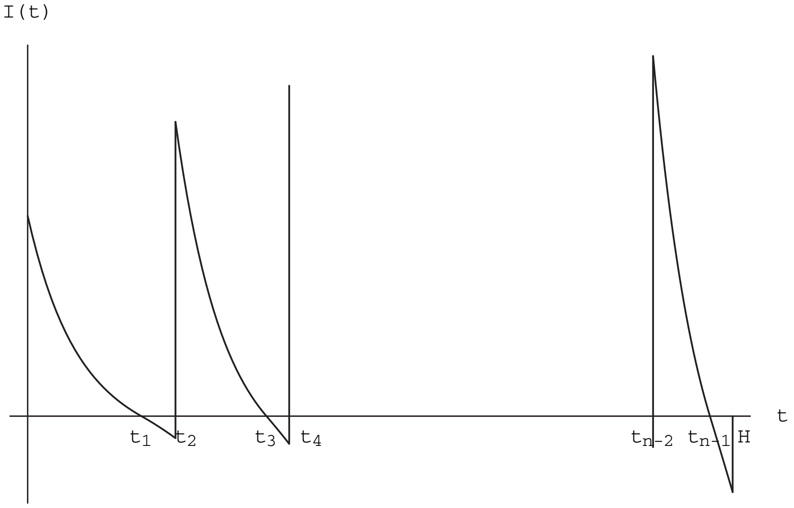


FIG. 3. *Typical inventory behavior with shortage for model CS.*

Theorem 6 establishes the existence of a unique optimal replenishment schedule. Theorem 7 states that replenishment periods within this schedule are successively strictly increasing if $\theta > 0$ and all of the same length if $\theta = 0$.

4. Extension to models with shortages. In this section, we assume that shortages are allowed. Existing models in the literature fall into one of the cases depicted in Figures 2–5. The difference between these models is whether or not there is a start with a period of shortage and whether or not there is an end with a period of shortage. To be specific, the cases are as follows:

(CC) *Start and end with a period of consumption* (Figure 2).

(CS) *Start with a period of consumption and end with a period of shortage* (Figure 3).

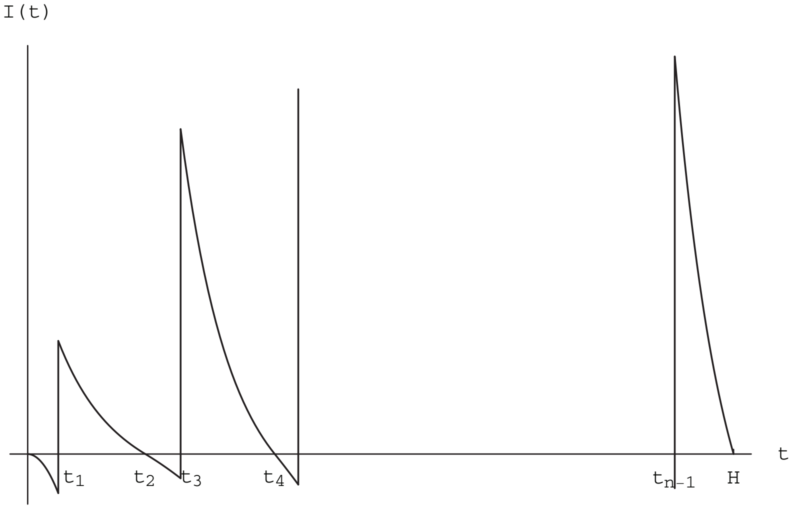


FIG. 4. *Typical inventory behavior with shortage for model SC.*

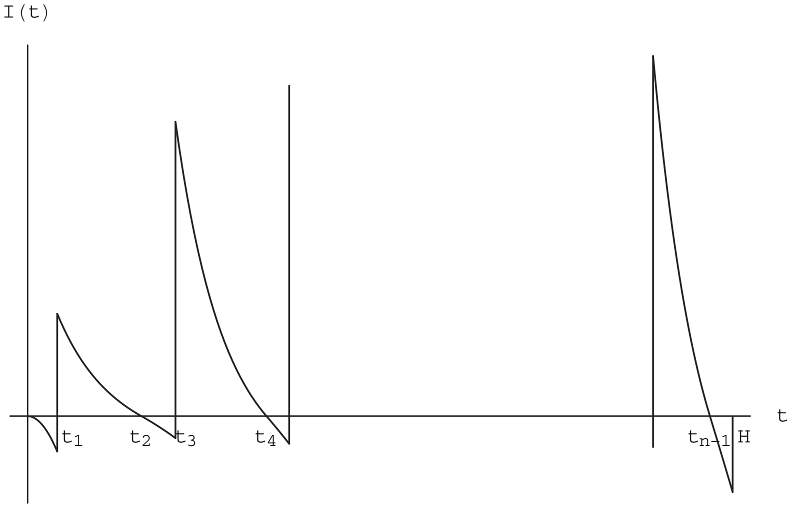


FIG. 5. *Typical inventory behavior with shortage for model SS.*

(SC) *Start with a period of shortage and end with a period of consumption* (Figure 4).

(SS) *Start and end with a period of shortage* (Figure 5).

In all of the existing models with shortages [3, 6, 14, 17, 18], the evolution of the inventory in a period of consumption has been described by (1.2) with θ constant, while the dynamics in a period of shortage have been given by

$$I'(t) = -D(t), \tag{4.1}$$

where $-I(t)$ denotes the level of shortage at time t.

We extend the description of the depletion of the inventory level during a period of consumption to (1.5) and in the interests of capturing as wide a range of compatible practical situations as feasible, we assume that the level of shortage in a shortage period is given by the complementary equation

$$I'(t) = -D_{\mathrm{s}}(t)F_{\mathrm{s}}\left(-e^{\Theta_{\mathrm{s}}(t)}I(t)\right) - \theta_{\mathrm{s}}(t)I(t), \tag{4.2}$$

where

$$\Theta_{\mathrm{s}}(t) = \int_0^t \theta_{\mathrm{s}}(u)\,du.$$

The assumptions on the various functions in this equation are largely analogous to those on the functions in (1.5), viz., $D_{\mathrm{s}} \in C([0,H]) \cap C^1(0,H)$ is positive on $(0,H)$, $\theta_{\mathrm{s}} \in C([0,H)) \cap C^1(0,H)$, and $F_{\mathrm{s}} \in C(0,\ell_{\mathrm{s}})$ is positive on $(0,\ell_{\mathrm{s}})$ for some number $\ell_{\mathrm{s}} > 0$ such that

$$\mu_{\mathrm{s}}(0,H) \leq \int_0^{\ell_{\mathrm{s}}} \frac{dv}{F_{\mathrm{s}}(v)} < \infty,$$

where

$$\mu_{\mathrm{s}}(x,y) = \int_x^y D_{\mathrm{s}}(t)\,e^{\Theta_{\mathrm{s}}(t)}\,dt.$$

Note that the functions in (4.2) need not be the same as in (1.5). Moreover, it is not assumed that θ_{s} is nonnegative. Thus, one can avoid thinking of the presence of the last term in the dynamics of (4.2) as a deterioration rate. The whole of the right-hand side of (4.2) should be interpreted as a single function describing the rate of change of the shortage level $-I$ in terms of time t and the shortage level itself. In this light, (4.2) is a broad generalization of (4.1).

Equation (4.2) holds in some time interval $t_{j-1} < t < t_j$, with the boundary condition

$$I(t) \to 0 \quad \text{as } t \downarrow t_{j-1}.$$

A solution is given by

$$I(t) = -\kappa_{\mathrm{s}}(t_{j-1}, t),$$

where

$$\kappa_{\mathrm{s}}(x,y) = e^{-\Theta_{\mathrm{s}}(y)}\psi_{\mathrm{s}}(\mu_{\mathrm{s}}(x,y)) \tag{4.3}$$

and ψ_{s} is defined by analogy to (1.9).

We shall further assume that during a period of shortage, there is a fixed cost per unit of the item per unit time c_3. When the evolution of the shortage level is as described above, the cost accumulated during the time-interval (t_{j-1}, t_j) following the OHD model is

$$c_3 \int_{t_{j-1}}^{t_j} |I(t)|\;dt = \int_{t_{j-1}}^{t_j} c_3\kappa_{\mathrm{s}}(t_{j-1}, t)\,dt. \tag{4.4}$$

Following the OHP model, a time-interval (t_{j-1}, t_j) of shortage leads to a cost of

$$c_2\,|I(t_j)| + c_3 \int_{t_{j-1}}^{t_j} |I(t)|\;dt = c_2\kappa_{\mathrm{s}}(t_{j-1}, t_j) + \int_{t_{j-1}}^{t_j} c_3\kappa_{\mathrm{s}}(t_{j-1}, t)\,dt. \tag{4.5}$$

For each of the different shortage models in the OHD case, we are subsequently led to a total cost function of the form (2.1), (2.2) with the following structure.

(CC) The total number of periods n is odd, $\nu_n = (n+1)/2$ and R_j is given by (1.20) for odd j and by

$$R_{\mathrm{s}}(x,y) = \int_x^y c_3 \kappa_{\mathrm{s}}(x,t)\,dt \tag{4.6}$$

for even j.

(CS) The total number of periods n is even, $\nu_n = n/2$, and R_j is as in the (CC) case.

(SC) The total number of periods n is even, $\nu_n = n/2$, and R_j is given by (4.6) for odd j and by (1.20) for even j.

(SS) The total number of periods n is odd, $\nu_n = (n-1)/2$, and R_j is as in the (SC) case.

In the OHP case, each of the different shortage models gives rise to a total cost function of the form (2.1), (2.2) with exactly the same structure. However, (1.20) is replaced by (1.21), and (4.6) is replaced by

$$R_{\mathrm{s}}(x,y) = c_2\kappa_{\mathrm{s}}(x,y) + \int_x^y c_3 \kappa_{\mathrm{s}}(x,t)\,dt. \tag{4.7}$$

The general theory developed in section 2 can subsequently be applied to all of these models if the respective functions $\{R_j\}_{j=1}^n$ are known to satisfy Hypotheses 1 and 2. Since we already know how (1.20) and (1.21) conform to these hypotheses, it remains to find the complementary results for (4.6) and (4.7).

4.1. The OHD cost function. The following lemmata are the respective analogies of Lemmata 4–7.

LEMMA 12. *Let R_{s} be given by* (4.6) *with κ_{s} given by* (4.3)*. Suppose that $F_{\mathrm{s}} \in C^1(0,\ell_{\mathrm{s}})$ or $D_{\mathrm{s}} = e^{-\Theta_{\mathrm{s}}}$. Then, R_{s} satisfies Hypothesis* 1*, and $\partial_x R_{\mathrm{s}} = \partial_y R_{\mathrm{s}} = 0$ on $\overline{\Omega}\setminus\Omega$.*

Proof. Without loss of generality, we may assume that $c_3 = 1$. Set

$$\rho_{\mathrm{s}}(x,y) = e^{\Theta_{\mathrm{s}}(x)-\Theta_{\mathrm{s}}(y)} \quad \text{and} \quad \gamma_{\mathrm{s}}(x,y) = e^{\Theta_{\mathrm{s}}(y)}\kappa_{\mathrm{s}}(x,y).$$

For the time being, we drop the subscript s without risk of confusion. We calculate

$$\partial_x\kappa = -D(x)\rho(x,y)F(\gamma(x,y))$$

and

$$\partial_y\kappa = D(y)F(\gamma(x,y)) - \theta(y)\kappa(x,y).$$

Further computation gives

$$\partial_x R = -D(x)Q(x,y),$$

where

$$Q(x,y) = \int_x^y \rho(x,t)F(\gamma(x,t))\,dt,$$

$$\partial_y R = \kappa(x,y),$$

$$\partial_x^2 R = -D'(x)Q(x,y) - D(x)\left(\partial_x Q\right)(x,y),$$

$$\partial_x\partial_y R = -D(x)\rho(x,y)F(\gamma(x,y)),$$

and

$$\partial_y^2 R = D(y)F(\gamma(x,y)) - \theta(y)\kappa(x,y).$$

Rewriting

$$Q(x,y) = \int_0^{y-x} \rho(x,x+u)F(\gamma(x,x+u))\,du,$$

when $F \in C^1(0,\ell)$, there holds

$$\begin{aligned}\partial_x Q = &-\rho(x,y)F(\gamma(x,y)) \\ &+ \int_0^{y-x} \rho(x,x+u)\{\theta(x)-\theta(x+u)\}F(\gamma(x,x+u))\,du \\ &+ \int_0^{y-x} \rho(x,x+u)\,(FF')\,(\gamma(x,x+u))q(x,x+u)\,du,\end{aligned}$$

where q is given by (3.4). Hence,

$$\begin{aligned}\partial_x Q = &-\rho(x,y)F(\gamma(x,y)) \\ &+ \int_x^y \{\theta(x)-\theta(t)\}\rho(x,t)F(\gamma(x,t))\,dt \\ &+ \int_x^y \rho(x,t)q(x,t)\,(FF')\,(\gamma(x,t))\,dt.\end{aligned}$$

As in the proof of Lemma 4, differentiation under the integral sign is justified under an appropriate change of variable. When $D = e^{-\Theta}$, it can be determined that $\gamma(x,x+u) = \psi(u)$, so the terms involving F' do not appear. □

LEMMA 13. *Suppose, in addition to the hypotheses of Lemma* 12, *that*

$$\Phi_{\rm s}(t) = \frac{D_{\rm s}'(t)}{D_{\rm s}(t)} + \theta_{\rm s}(t)$$

is nonincreasing on $(0,H)$. *Then, if* $D_{\rm s} = e^{-\Theta_{\rm s}}$ *or* $F_{\rm s}(v) = v^{\beta_{\rm s}}$ *for some* $\beta_{\rm s} < 1$, *there holds* $\mathcal{L}_y R_{\rm s} \geq 0$ *in* Ω, *where*

$$f(t) = \theta_{\rm s}(t) \tag{4.8}$$

or

$$f(t) = \theta_{\rm s}(t) - \frac{\Phi_{\rm s}(t)}{1-\beta_{\rm s}}, \tag{4.9}$$

respectively.

Proof. Continuing from the proof of Lemma 12, it can be deduced that for any operator of the form (2.10), there holds

$$\mathcal{L}_y R = \kappa(x,y)\phi(x,y),$$

where

$$\phi(x,y) = f(y) - \theta(y) + q(x,y)\frac{F(\gamma(x,y))}{\gamma(x,y))}.$$

If Φ is nonincreasing, Lemma 1 implies that

$$\phi(x,y) \geq f(y) - \theta(y) + \Phi(y)\frac{\Psi(v)F(v)}{v},$$

where Ψ is defined by (3.5) and $v = \gamma(x,y)$. The asserted follows from Lemmata 2 and 3. □

LEMMA 14. *Suppose, in addition to the hypotheses of Lemma* 13, *that* θ_{s} *is nondecreasing on* $(0,H)$. *Then, the conclusions of, Lemma* 13 *apply to* $\mathcal{L}_x R_{\mathrm{s}}$ *as well as* $\mathcal{L}_y R_{\mathrm{s}}$.

Proof. Further to the proof of the previous two lemmata, for any operator of the form (2.9), there holds

$$\mathcal{L}_x R = D(x)\int_x^y \rho(x,t)F(\gamma(x,t))\phi(x,t)\,dt,$$

where

$$\phi(x,y) = \theta(y) - \Phi(x) - q(x,y)F'(\gamma(x,y)) - f(x)$$

when $F \in C^1(0,\ell)$. Now, if θ is nondecreasing and Φ is nonincreasing,

$$\phi(x,y) \geq \theta(x) - \Phi(x)\left\{1 + \Psi(v)F'(v)\right\} - f(x),$$

where $v = \gamma(x,y)$. The result follows via Lemmata 2 and 3. When $D = e^{-\Theta}$, there holds $\phi \equiv 0$, and the terms involving F' can be ignored. □

LEMMA 15. *Let* R_{s} *be given by* (4.6) *with* $c_3 = 1$ *and* $(x,y) \in \Omega$.

(i) *Suppose that* $D_{\mathrm{s}} = e^{-\Theta_{\mathrm{s}}}$. *Then,*

$$(\partial_y R_{\mathrm{s}})(x,y) = D_{\mathrm{s}}(y)\psi_{\mathrm{s}}(y-x). \tag{4.10}$$

Furthermore, if D_{s} *is nondecreasing on* $(0,H)$, *then*

$$(-\partial_x R_{\mathrm{s}})(x,y) \geq D_{\mathrm{s}}(x)\psi_{\mathrm{s}}(y-x), \tag{4.11}$$

with equality if and only if D_{s} *is constant on* (x,y). *On the other hand, if* D_{s} *is nonincreasing on* $(0,H)$, *then* (4.11) *holds with the inequality sign reversed, and equality as stated before.*

(ii) *Suppose that* $F_{\mathrm{s}}(v) = v^{\beta_{\mathrm{s}}}$ *for some* $0 \leq \beta_{\mathrm{s}} < 1$. *Let*

$$A_{\mathrm{s}}(t) = \left\{D_{\mathrm{s}}(t)\,e^{\beta_{\mathrm{s}}\Theta_{\mathrm{s}}(t)}\right\}^{1/(1-\beta_{\mathrm{s}})}. \tag{4.12}$$

If there exists a number λ_{s} *such that* $D_{\mathrm{s}}e^{\lambda_{\mathrm{s}}\Theta_{\mathrm{s}}}$, $(\beta_{\mathrm{s}} - \lambda_{\mathrm{s}})\Theta_{\mathrm{s}}$, *and* $-(1-\lambda_{\mathrm{s}})\theta_{\mathrm{s}}$ *are nondecreasing on* $(0,H)$, *then*

$$(\partial_y R_{\mathrm{s}})(x,y) \leq A_{\mathrm{s}}(y)\psi_{\mathrm{s}}\left(\int_0^{y-x} e^{-(1-\lambda_{\mathrm{s}})\theta_{\mathrm{s}}(y)u}\,du\right) \tag{4.13}$$

and

$$(-\partial_x R_{\mathrm{s}})(x,y) \geq A_{\mathrm{s}}(x)\psi_{\mathrm{s}}\left(\int_0^{y-x} e^{-(1-\lambda_{\mathrm{s}})\theta_{\mathrm{s}}(x)u}\,du\right), \tag{4.14}$$

with equality in both simultaneously if and only if $A_{\mathrm{s}}^{1-\beta_{\mathrm{s}}} - (1-\lambda_{\mathrm{s}})\theta_{\mathrm{s}}$ *is constant on* (x,y). *On the other hand, if there exists a number* λ_{s} *such that the aforesaid functions are nonincreasing on* $(0,H)$, *then* (4.13) *and* (4.14) *hold with the inequality sign reversed, and equality in both simultaneously as stated before.*

Proof. Dropping the subscripts s and imitating the start of the proof of Lemma 7, we find

$$(\partial_y R)(x,y) = e^{-\Theta(y)}\psi(\mu(x,y)) \tag{4.15}$$

and

$$(-\partial_x R)(x,y) = D(x)\int_x^y \rho(x,t)\psi'(\mu(x,t))\,dt. \tag{4.16}$$

Hence, if $D = e^{-\Theta}$, we have (4.10), and

$$(-\partial_x R)(x,y) = \int_x^y D(t)\psi'(t-x)\,dt.$$

The latter gives (4.11) if D is nondecreasing with equality if and only if $D(x) = D(y)$. On the other hand, if D is nonincreasing, it gives (4.11) with equality as stated. This proves part (i). Suppose next that ψ is given by (1.12) for some $0 \le \beta < 1$ and $De^{\lambda\Theta}$ is nondecreasing for some number λ. Then

$$\begin{aligned} D(x)\,e^{\Theta(x)}\rho^{\lambda-1}(x,t)\int_x^t \rho^{1-\lambda}(u,t)\,du &\le \mu(x,t) \\ &\le D(t)\,e^{\Theta(t)}\int_x^t \rho^{1-\lambda}(u,t)\,du \end{aligned} \tag{4.17}$$

for all $x < t \le y$. Hence, (4.15) gives

$$(\partial_y R)(x,y) \le A(y)\psi\left(\textstyle\int_x^y \rho^{1-\lambda}(t,y)\,dt\right),$$

while (4.16) implies

$$(-\partial_x R)(x,y) \ge A(x)\int_x^y \rho^{1+\beta(\lambda-1)/(1-\beta)}(x,t)\psi'\left(\textstyle\int_x^t \rho^{1-\lambda}(x,u)\,du\right)\,dt.$$

Subsequently, if $(\beta-\lambda)\Theta$ and $-(1-\lambda)\theta$ are nondecreasing, we obtain (4.13) and (4.14). Moreover, retracing the steps, it can be seen that the former holds with equality if and only if $D\,e^{\lambda\Theta}$ and $-(1-\lambda)\theta$ are constant on (x,y), while the latter holds with equality if $De^{\lambda\Theta}$, $(\beta-\lambda)\Theta$, and $-(1-\lambda)\theta$ are constant on (x,y) and only if $(\beta-\lambda)\Theta$ and $-(1-\lambda)\theta$ are constant on (x,y). The proof of the reversed inequalities is entirely analogous. □

Amalgamating Lemmata 4–6 and 12–14 with Theorems 1 and 2 yields the following regarding the OHD model with shortages.

THEOREM 8. *Consider model* (2.1), (2.2) *where R_j is given alternately by* (1.20) *under the introductory assumptions of section* 1 *and* (4.6) *under the introductory assumptions of this section. Suppose, furthermore, that $D'/D+\theta$, $w'/w-\theta$, $D_{\mathrm{s}}'/D_{\mathrm{s}}+\theta_{\mathrm{s}}$, and $-\theta_{\mathrm{s}}$ are nonincreasing on $(0,H)$. Then, for every $n \ge 1$, S_n has a unique minimum with respect to $t_0,\ldots,t_n$ satisfying* (1.1) *when any of the following hold:*

(a) *$D = e^{-\Theta}$, $D_{\mathrm{s}} = e^{-\Theta_{\mathrm{s}}}$, and*

$$\theta_{\mathrm{s}} = \theta - \frac{w'}{w}; \tag{4.18}$$

(b) $D = e^{-\Theta}$, $F_{\rm s}(v) = v^{\beta_{\rm s}}$ *for some* $\beta_{\rm s} < 1$, *and*

$$\frac{1}{1-\beta_{\rm s}}\left(\frac{D_{\rm s}'}{D_{\rm s}} + \beta_{\rm s}\theta_{\rm s}\right) = \frac{w'}{w} - \theta; \tag{4.19}$$

(c) $F(v) = v^{\beta}$ *for some* $\beta < 1$, $D_{\rm s} = e^{-\Theta_{\rm s}}$, *and*

$$\theta_{\rm s} = -\frac{1}{1-\beta}\left(\frac{D'}{D} + \beta\theta\right) - \frac{w'}{w};$$

(d) $F(v) = v^{\beta}$ *for some* $\beta < 1$, $F_{\rm s}(v) = v^{\beta_{\rm s}}$ *for some* $\beta_{\rm s} < 1$, *and*

$$\frac{1}{1-\beta_{\rm s}}\left(\frac{D_{\rm s}'}{D_{\rm s}} + \beta_{\rm s}\theta_{\rm s}\right) = \frac{1}{1-\beta}\left(\frac{D'}{D} + \beta\theta\right) + \frac{w'}{w}.$$

Moreover, if s_n *denotes the minimum value of* S_n *under constraint* (1.1), *then* s_n *is a strictly decreasing function of* $n \geq 1$, *and* s_{2i} *and* s_{2i-1} *are strictly convex functions of* $i \geq 1$.

With the convention that $x|_{\rm XX}$ denotes the value of the variable x associated with the shortage model XX and XX stands for any one of the combinations CC, CS, SC, or SS, we have the following result.

COROLLARY 8.1. *For any* $i \geq 1$, *there holds*

$$\nu_{2i-1}|_{\rm CC} = \nu_{2i}|_{\rm CS} = \nu_{2i}|_{\rm SC} = \nu_{2i+1}|_{\rm SS} = i$$

and

$$s_{2i-1}|_{\rm CC}\,,\, s_{2i-1}|_{\rm SS} > s_{2i}|_{\rm CS}\,,\, s_{2i}|_{\rm SC} > s_{2i+1}|_{\rm CC}\,,\, s_{2i+1}|_{\rm SS}\,.$$

COROLLARY 8.2. *The following alternatives are mutually exclusive for the models CC and SS.*

(i) *If* $K > s_1 - s_3$, *then* $\mathcal{C}$ *has a unique minimum which occurs for* $n = 1$.

(ii) *If there exists a number* $N \geq 2$ *such that* $s_{2N-3} - s_{2N-1} > K > s_{2N-1} - s_{2N+1}$, *then* $\mathcal{C}$ *has a unique minimum which occurs for* $n = 2N - 1$.

(iii) *If there exists a number* $N \geq 1$ *such that* $K = s_{2N-1} - s_{2N+1}$, *then* $\mathcal{C}$ *has precisely two minima which occur for* $n = 2N - 1$ *and* $n = 2N + 1$.

Whereas, the following alternatives are mutually exclusive for the models CS and SC.

(i) *If* $K > s_2 - s_4$, *then* $\mathcal{C}$ *has a unique minimum which occurs for* $n = 2$.

(ii) *If there exists a number* $N \geq 2$ *such that* $s_{2N-2} - s_{2N} > K > s_{2N} - s_{2N+2}$, *then* $\mathcal{C}$ *has a unique minimum which occurs for* $n = 2N$.

(iii) *If there exists a number* $N \geq 1$ *such that* $K = s_{2N} - s_{2N+2}$, *then* $\mathcal{C}$ *has precisely two minima which occur for* $n = 2N$ *and* $n = 2N + 2$.

THEOREM 9. *Let* $t_0, t_1, \ldots, t_n$ *be the minimum of* S_n *under constraint* (1.1) *given by Theorem* 8 *for some* $n \geq 3$ *under one of the following conditions:*

(a) $D = e^{-\Theta}$ *and* $D_{\rm s} = e^{-\Theta_{\rm s}}$;

(b) $D = e^{-\Theta}$ *and* $F_{\rm s}(v) = v^{\beta_{\rm s}}$ *for some* $0 \leq \beta_{\rm s} < 1$;

(c) $F(v) = v^{\beta}$ *for some* $0 \leq \beta < 1$ *and* $D_{\rm s} = e^{-\Theta_{\rm s}}$;

(d) $F(v) = v^{\beta}$ *for some* $0 \leq \beta < 1$ *and* $F_{\rm s}(v) = v^{\beta_{\rm s}}$ *for some* $0 \leq \beta_{\rm s} < 1$.

Then, in each case, respectively, the length of two consecutive consumption periods and the length of two consecutive shortage periods decreases when the following functions are nondecreasing on $(0, H)$:

(a) wD;

(b) $D_{\rm s}e^{\lambda_{\rm s}\Theta_{\rm s}}$, $(\beta_{\rm s} - \lambda_{\rm s})\Theta_{\rm s}$, *and* $-(1 - \lambda_{\rm s})\theta_{\rm s}$ *for some number* $\lambda_{\rm s}$;

(c) *$De^{\lambda\Theta}$, $we^{\{(\beta-\lambda)/(1-\beta)\}\Theta}$, and $(1-\lambda)\theta$ for some number λ;*

(d) *$De^{\lambda\Theta}$, $we^{\{(\beta-\lambda)/(1-\beta)\}\Theta}$, and $(1-\lambda)\theta$ for some number λ, and $D_{\rm s}e^{\lambda_{\rm s}\Theta_{\rm s}}$, $(\beta_{\rm s}-\lambda_{\rm s})\Theta_{\rm s}$, and $-(1-\lambda_{\rm s})\theta_{\rm s}$ for some number $\lambda_{\rm s}$.*

Moreover, in each case, the length of two consecutive consumption periods and the length of two consecutive shortage periods is strictly decreasing if one of the following functions is strictly increasing on $(0,H)$:

(a) *wD;*

(b) *wD or $D_{\rm s}e^{\beta_{\rm s}\Theta_{\rm s}}-(1-\lambda_{\rm s})\theta_{\rm s}$;*

(c) *$Dw^{1-\beta}e^{\beta\Theta}+(1-\lambda)\theta$;*

(d) *$Dw^{1-\beta}e^{\beta\Theta}+(1-\lambda)\theta$ or $D_{\rm s}e^{\beta_{\rm s}\Theta_{\rm s}}-(1-\lambda_{\rm s})\theta_{\rm s}$.*

On the other hand, in each of the initial cases, the length of two consecutive consumption periods and the length of two consecutive shortage periods increases when the functions in the subsequent list are nonincreasing on $(0,H)$. Moreover, in each case, respectively, the length of two consecutive consumption periods and the length of two consecutive shortage periods is strictly increasing if one of the functions in the final list is strictly decreasing on $(0,H)$.

Proof. According to Theorem 3, the length of consecutive consumption periods decreases if

$$\frac{(\partial_y R)\,(x-\delta,x)}{(-\partial_x R)\,(y,y+\delta)} \le \frac{(-\partial_x R_{\rm s})\,(x,y)}{(\partial_y R_{\rm s})\,(x,y)} \tag{4.20}$$

for all $0\le x-\delta<x<y<y+\delta\le H$. Moreover, the length of these periods decreases strictly if (4.20) is strict in all such circumstances. Likewise, the length of consecutive shortage periods decreases if

$$\frac{(\partial_y R_{\rm s})\,(x-\delta,x)}{(-\partial_x R_{\rm s})\,(y,y+\delta)} \le \frac{(-\partial_x R)\,(x,y)}{(\partial_y R)\,(x,y)}. \tag{4.21}$$

Moreover, the length of these periods decreases strictly if (4.21) is strict for all $0\le x-\delta<x<y<y+\delta\le H$. Sufficient conditions under which (4.20) and (4.21) hold can be found by employing Lemmata 7 and 15. For instance, in case (a), conditions can be found by assuming that wD and $D_{\rm s}$ are nondecreasing. Substituting (3.15), (3.16), (4.10), and (4.11) in (4.20) and (4.21) leads to the conditions

$$\frac{w(x)D(x)\psi(\delta)}{w(y)D(y)\psi(\delta)} \le \frac{D_{\rm s}(x)\psi_{\rm s}(y-x)}{D_{\rm s}(y)\psi_{\rm s}(y-x)}$$

and

$$\frac{D_{\rm s}(x)\psi_{\rm s}(\delta)}{D_{\rm s}(y)\psi_{\rm s}(\delta)} \le \frac{w(x)D(x)\psi(y-x)}{w(y)D(y)\psi(y-x)},$$

respectively. Noting that (4.18) implies that $D_{\rm s}/wD$ is constant on $(0,H)$, this gives the first assertion related to case (a) in the theorem. Furthermore, it gives (4.20) and (4.21) with strict inequality if wD is strictly monotonic. In case (b), assuming that wD is nondecreasing, there exists a number $\lambda_{\rm s}$ such that $D_{\rm s}e^{\lambda_{\rm s}\Theta_{\rm s}}$, $(\beta_{\rm s}-\lambda_{\rm s})\Theta_{\rm s}$, and $-(1-\lambda_{\rm s})\theta_{\rm s}$ are nondecreasing, substituting (3.15), (3.16), (4.13), and (4.14) in (4.20) and (4.21) leads to the identification of

$$\frac{w(x)D(x)\psi(\delta)}{w(y)D(y)\psi(\delta)} \le \frac{A_{\rm s}(x)\psi_{\rm s}\left(\int_0^{y-x} e^{-(1-\lambda_{\rm s})\theta_{\rm s}(x)u}\,du\right)}{A_{\rm s}(y)\psi_{\rm s}\left(\int_0^{y-x} e^{-(1-\lambda_{\rm s})\theta_{\rm s}(y)u}\,du\right)}$$

and

$$\frac{A_s(x)\psi_s\left(\int_0^\delta e^{-(1-\lambda_s)\theta_s(x)u}\,du\right)}{A_s(y)\psi_s\left(\int_0^\delta e^{-(1-\lambda_s)\theta_s(y)u}\,du\right)} \le \frac{w(x)D(x)\psi(y-x)}{w(y)D(y)\psi(y-x)}$$

as sufficient conditions for the monotonicity of the length of the consumption and shortage periods, respectively. Since (4.19) implies that A_s/wD is constant on $(0,H)$, this gives the first assertion of the theorem regarding case (b). The further assertion of strict monotonicity follows from the conditions given by Lemmata 7 and 15 under which (3.15), (3.16), (4.13), and (4.14) may be strict. The proofs of the assertions concerning increment of the length of consumption and shortage periods in cases (c) and (d) are similar. With regard to the remainder of the theorem, we note that, by Theorem 3, the length of two consecutive consumption periods increases if (4.20) holds with the inequality reversed for all $0 \le x-\delta < x < y < y+\delta \le H$. Likewise, the length of two consecutive consumption periods increases if (4.21) holds with the inequality reversed. Moreover, the length of such periods is strictly monotonic if the appropriate inequality is strict. Retracing the proof so far leads to the desired results. □

COROLLARY 9.1. *In each of the respective cases in Theorem* 9, *if the functions in the second list are constant on* $(0,H)$, *then all consumption periods have the same length and all shortage periods have the same length.*

Using Theorem 3 and Lemmata 7 and 15, it is actually possible to identify some special circumstances in which the length of adjacent periods is monotonic irrespective of whether they specifically pertain to consumption or shortage periods.

COROLLARY 9.2. *Suppose that* $F_s = F$ *and* $c_3D_s = wD$ *in case* (a) *of Theorem* 9, *or* $\beta_s = \beta$, $(1-\lambda_s)\theta_s = -(1-\lambda)\theta$, *and* $c_3(D_s e^{\beta\Theta_s})^{1/(1-\beta)} = w(De^{\beta\Theta})^{1/(1-\beta)}$ *in case* (d) *of Theorem* 9. *Then, if the respective functions in the second list of Theorem* 9 *are nondecreasing on* $(0,H)$, (3.28) *holds. On the other hand, if these functions are nonincreasing on* $(0,H)$, (3.29) *holds. Consequently, if these functions are constant on* $(0,H)$, (3.30) *applies.*

The results previously known for the existence of an optimal replenishment schedule for a finite time horizon inventory model with shortages [3, 4, 6] are readily recovered from Theorem 8. Taking $F_s \equiv F \equiv 1$ in part (d), the theorem establishes the existence of an optimal schedule when the consumption dynamics are described by (1.2) and the shortage dynamics by

$$I'(t) = -\alpha w(t)D(t) - \theta_s(t)I(t)$$

for some number $\alpha > 0$, under the hypotheses on D, θ, and θ_s assumed throughout this paper plus the supplementary hypothesis that $D'/D+\theta$, $w'/w-\theta$, $(wD)'/wD+\theta_s$, and $-\theta_s$ are nonincreasing. Note that the monotonicity of the first two functions in the supplementary hypothesis is precisely the sufficient condition for the existence of an optimal replenishment schedule for the OHD cost function without shortages (stated in Theorem 4). Consequently, the monotonicity of the other functions may be viewed purely as a restriction on θ_s. In the event that θ is constant, w is constant too. Subsequently, taking $\alpha = 1/w$ and $\theta_s \equiv 0$, we obtain the existence of an optimal schedule for consumption model (1.2) and shortage model (4.1), under the single hypothesis over and above those assumed throughout that D'/D is nonincreasing. All previous results [3, 4, 6, 14] have been obtained in these circumstances. Indeed, they have required the additional assumption that $D' \in C([0,H])$ and $D'(t) \neq 0$

for $0 \leq t \leq H$. When $D'(t) > 0$ for $0 \leq t \leq H$, the length of two consecutive consumption periods and the length of two consecutive shortage periods in the optimal replenishment schedule is strictly decreasing. When $D'(t) < 0$ for $0 \leq t \leq H$, these lengths are strictly increasing. These results can be recovered from Theorem 9 by letting $\beta = \beta_{\mathrm{s}} = \lambda = \lambda_{\mathrm{s}} = 0$ in part (d). In their generalization, Theorems 8 and 9 plainly cover a host of alternatives to the previously considered combination of the models (1.2) and (4.1), with or without the assumption that θ is constant and the ensuing hypotheses.

4.2. The OHP cost function. Given that function (4.6) can be put into the framework developed in section 2, to establish under which conditions the function (4.7) is amenable to the same treatment, it suffices to examine the influence of the extra term. This is given by the following four lemmata. These are the analogies of Lemmata 8–11, respectively.

Lemma 16. *Let κ_{s} be given by* (4.3). *Suppose that $\theta_{\mathrm{s}} \leq 0$, and $F_{\mathrm{s}} \in C([0,\ell_{\mathrm{s}}]) \cap C^1(0,\ell_{\mathrm{s}})$ is nondecreasing on $(0,H)$. Then, κ_{s} satisfies Hypothesis* 1 *with weak inequality in* (2.6), *and $\partial_y \kappa_{\mathrm{s}} = -\partial_x \kappa_{\mathrm{s}} = D_{\mathrm{s}}(x) F_{\mathrm{s}}(0)$ on $\overline{\Omega} \setminus \Omega$.*

Proof. Carrying on from the proof of Lemma 12 with the subscript s dropped from the notation, one can compute

$$\partial_x^2 \kappa = -\{D'(x) + D(x)\theta(x)\}\rho(x,y)F(\gamma(x,y)) + D^2(x)\, e^{\Theta(x)} \rho(x,y)\,(FF')\,(\gamma(x,y)),$$

$$\partial_x \partial_y \kappa = D(x)\theta(y)\rho(x,y)F(\gamma(x,y)) - D(x)D(y)\, e^{\Theta(x)}\,(FF')\,(\gamma(x,y)),$$

and

$$\begin{aligned}\partial_y^2 \kappa = {} & \{D'(y) - D(y)\theta(y)\}F(\gamma(x,y)) \\ & + D^2(y)\, e^{\Theta(y)}\,(FF')\,(\gamma(x,y)) + \left\{\theta^2(y) - \theta'(y)\right\}\kappa(x,y).\end{aligned}$$

The lemma follows from these expressions and those established for $\partial_x \kappa$ and $\partial_y \kappa$ in the proof of Lemma 12. □

Lemma 17. *Suppose, in addition to the hypotheses of Lemma* 16, *that Φ_{s} is nonincreasing, and θ_{s} is nondecreasing on $(0,H)$. Then, if $D_{\mathrm{s}} = e^{-\Theta_{\mathrm{s}}}$ or $F_{\mathrm{s}}(v) = v^{\beta_{\mathrm{s}}}$ for some $0 \leq \beta_{\mathrm{s}} < 1$, there holds $\mathcal{L}_x \kappa_{\mathrm{s}} \geq 0$ in Ω, where f is given by* (4.8) *or* (4.9), *respectively.*

Proof. Yet again ignoring the subscript s and employing the calculations above, one finds

$$\mathcal{L}_x \kappa = D(x)\rho(x,y)F(v)\phi(x,y)$$

for any operator of the form (2.9), where

$$\phi(x,y) = \theta(y) - \Phi(x) - q(x,y)F'(v) - f(x)$$

and $v = \gamma(x,y)$. Recalling Lemma 1, when Φ is nonincreasing and θ is nondecreasing,

$$\phi(x,y) \geq \theta(x) - \Phi(x)\left\{1 + \Psi(v)F'(v)\right\} - f(x).$$

This implies $\mathcal{L}_x \kappa \geq 0$ under the conditions stated, via Lemmata 2 and 3. □

Lemma 18. *Suppose, in addition to the hypotheses of Lemma* 16, *that Φ_{s} and θ_{s} are nonincreasing on $(0,H)$. Then, the conclusions of the previous lemma apply with $\mathcal{L}_y \kappa_{\mathrm{s}}$ in lieu of $\mathcal{L}_x \kappa_{\mathrm{s}}$.*

Proof. Substituting the requisite expressions in (2.10), one finds

$$\mathcal{L}_y\kappa = D(y)F(v)\phi_1(x,y) + \kappa(x,y)\phi_2(x,y),$$

where

$$\phi_1(x,y) = f(y) - \theta(y) + \Phi(y) + q(x,y)F'(v),$$

$$\phi_2(x,y) = \theta^2(y) - \theta'(y) - \theta(y)q(x,y)\frac{F(v)}{v} - \theta(y)f(y),$$

and $v = \gamma(x,y)$. By Lemma 1, if Φ is nonincreasing, then

$$\phi_1(x,y) \geq f(y) - \theta(y) + \Phi(y)\left\{1 + \Psi(v)F'(v)\right\}.$$

Similarly,

$$\phi_2(x,y) \geq \theta^2(y) - \theta'(y) - \theta(y)\Phi(y)\frac{\Psi(v)F(v)}{v} - \theta(y)f(y).$$

Armed with these inequalities, the proof may be completed along the lines of that of Lemma 10. □

LEMMA 19. *Let the assumptions of Lemma* 16 *hold, and* $(x,y) \in \Omega$.

(i) *Suppose that* $D_{\mathrm{s}} = e^{-\Theta_{\mathrm{s}}}$. *If* θ_{s} *is nonincreasing on* $(0,H)$, *then*

$$(-\partial_x\kappa_{\mathrm{s}})(x,y) \geq D_{\mathrm{s}}(x)\{\psi_{\mathrm{s}}'(y-x) - \theta_{\mathrm{s}}(x)\psi_{\mathrm{s}}(y-x)\}. \tag{4.22}$$

Anyhow,

$$(\partial_y\kappa_{\mathrm{s}})(x,y) = D_{\mathrm{s}}(y)\{\psi_{\mathrm{s}}'(y-x) - \theta_{\mathrm{s}}(y)\psi_{\mathrm{s}}(y-x)\}. \tag{4.23}$$

(ii) *Suppose that* $F_{\mathrm{s}}(v) = v^{\beta_{\mathrm{s}}}$ *for some* $0 \leq \beta_{\mathrm{s}} < 1$. *Let* $A_{\mathrm{s}}(t)$ *be given by* (4.12). *If* $D_{\mathrm{s}}e^{\beta_{\mathrm{s}}\Theta_{\mathrm{s}}}$ *is nondecreasing and* θ_{s} *is nonincreasing on* $(0,H)$, *then*

$$(-\partial_x\kappa_{\mathrm{s}})(x,y) \geq A_{\mathrm{s}}(x)\, e^{-(1-\beta_{\mathrm{s}})\theta_{\mathrm{s}}(x)(y-x)}\psi_{\mathrm{s}}'\left(\int_0^{y-x} e^{-(1-\beta_{\mathrm{s}})\theta_{\mathrm{s}}(x)u}\,du\right) \tag{4.24}$$

and

$$(\partial_y\kappa_{\mathrm{s}})(x,y) \leq A_{\mathrm{s}}(y)\, e^{-(1-\beta_{\mathrm{s}})\theta_{\mathrm{s}}(y)(y-x)}\psi'\left(\int_0^{y-x} e^{-(1-\beta_{\mathrm{s}})\theta_{\mathrm{s}}(y)u}\,du\right). \tag{4.25}$$

On the other hand, if $D_{\mathrm{s}}e^{\beta_{\mathrm{s}}\Theta_{\mathrm{s}}}$ *is nonincreasing and* θ_{s} *is nondecreasing on* $(0,H)$, *then* (4.24) *and* (4.25) *hold with the inequality sign reversed.*

Proof. Forgetting about the subscript s, analogously to in the proof of Lemma 11, it can be verified that

$$(-\partial_x\kappa)(x,y) = D(x)\rho(x,y)\psi'(\mu(x,y)) \tag{4.26}$$

and

$$(\partial_y\kappa)(x,y) = D(y)\psi'(\mu(x,y)) - \theta(y)\, e^{-\Theta(y)}\psi(\mu(x,y)). \tag{4.27}$$

When $D = e^{-\Theta}$ and θ is nonincreasing, (4.26) implies

$$(-\partial_x\kappa)(x,y) \geq D(x)\, e^{-\theta(x)(y-x)}\psi'(y-x).$$

Consequently, to verify (4.22), it suffices to show that

$$e^{-\theta(x)(y-x)}\psi'(y-x) \geq \psi'(y-x) - \theta(x)\psi(y-x). \tag{4.28}$$

However, if F is nondecreasing on $[0,\ell]$, then $F(v)\Psi(v) \geq v$ for all $0 < v \leq \ell$, by (3.5). Hence, because $e^{\eta} \geq 1+\eta$ for any number η, there holds

$$F(v)\, e^{-\theta(x)\Psi(v)} \geq F(v)\{1-\theta(x)\Psi(v)\} \geq F(v) - \theta(x)v.$$

Substituting $v = \psi(y-x)$ in the above inequality and recalling (1.9) and (3.5) gives (4.28), and therewith, (4.22). When $D = e^{-\Theta}$, (4.27) reduces to (4.23). This proves part (i). To prove part (ii), we note that if $De^{\beta\Theta}$ is nondecreasing on $(0,H)$, then (4.17) with $\lambda = \beta$ holds for all $x \leq t < y$. Hence,

$$\left(-\partial_x\kappa\right)(x,y) \geq A(x)\rho^{1-\beta}(x,y)\psi'\left(\textstyle\int_x^y \rho^{1-\beta}(t,y)\,dt\right)$$

and

$$\left(\partial_y\kappa\right)(x,y) \leq A(y)\left\{\psi'\left(\textstyle\int_x^y \rho^{1-\beta}(t,y)\,dt\right) - \theta(y)\psi\left(\textstyle\int_x^y \rho^{1-\beta}(t,y)\,dt\right)\right\}.$$

If θ is nonincreasing, these inequalities yield (4.24) and (4.25). This confirms the first assertion of part (ii). Confirmation of the second is a replica of the first. □

Collating the above lemmata with those in section 3 and the previous subsection and applying Theorems 1–3 leads to the following results for the OHP model incorporating shortages.

THEOREM 10. *Consider model* (2.1), (2.2), *where* R_j *is given alternately by* (1.21) *under the introductory assumptions of section* 1 *and* (4.7) *under the introductory assumptions of this section. Suppose, furthermore, that* D'/D *and* $D_{\mathrm{s}}'/D_{\mathrm{s}}$ *are nonincreasing,* $\theta \geq 0$ *and* $\theta_{\mathrm{s}} \leq 0$ *are constant on* $(0,H)$, $F \in C([0,\ell]) \cap C^1(0,\ell)$ *is nondecreasing,* $F_{\mathrm{s}} \in C([0,\ell_{\mathrm{s}}]) \cap C^1(0,\ell_{\mathrm{s}})$ *is nondecreasing, and* $DF(0) \equiv D_{\mathrm{s}}F_{\mathrm{s}}(0)$ *on* $[0,H]$. *Then, for every* $n \geq 1$, S_n *has a unique minimum with respect to* $t_0,\ldots,t_n$ *satisfying* (1.1) *when any of the alternatives* (a), (b), (c), *or* (d) *of Theorem* 8 *hold with* $w \equiv c_1$. *Moreover, if* s_n *denotes the minimum value of* S_n *under the constraint* (1.1), *then* s_n *is a strictly decreasing function of* $n \geq 1$, *and* s_{2i} *and* s_{2i-1} *are strictly convex functions of* $i \geq 1$.

COROLLARY 10.1. *Verbatim, the conclusions of Corollaries* 8.1 *and* 8.2 *hold.*

THEOREM 11. *The conclusions of Theorem* 9, *Corollary* 9.1, *and Corollary* 9.2 *hold under the restriction that* $w \equiv c_1$, $\lambda = \beta$, *and* $\lambda_{\mathrm{s}} = \beta_{\mathrm{s}}$.

No prior results for the OHP model with shortages are known.

5. Conclusion.

In this paper, we have examined a class of optimization problems that arise as a result of a search for the optimal replenishment schedule for finite time horizon inventory models. The decision variables consist of a single integer variable and an ordered bounded set of nonnegative real numbers, whereby the cardinality of this set is equal to the value of the integer variable. It has been shown that if the objective function possesses certain generic properties and satisfies some partial differential inequalities—Hypotheses 1 and 2, respectively—then a solution to the optimization problem exists and is unique. The hypotheses are satisfied by existing models in the literature and a range of new models. Special properties of the optimal solution have also been presented.

The tool used for tackling the finite time horizon inventory model is new in nature and can easily be adapted to deal with models with partial backlogging. This is because the cost function $\mathcal{C}$ retains the separable structure (2.1), (2.2) with functions R_j that satisfy Hypothesis 1.

The inclusion of inflation in the model, as proposed in [18], requires a certain adaptation. In place of (1.13), the OHD cost function for the model without shortages becomes

$$\mathcal{C} = \sum_{j=1}^{n} \left\{ K e^{-r_{\mathrm{i}} t_{j-1}} + \int_{t_{j-1}}^{t_j} c_1(t) I(t)\, dt + c_2 e^{-r_{\mathrm{e}} t_{j-1}} \int_{t_{j-1}}^{t_j} \theta(t) I(t)\, dt \right\}, \tag{5.1}$$

where r_{i} denotes a net discount rate accounting for internal inflation, r_{e} denotes a net discount rate accounting for external inflation

$$c_1(t) = c_{1,\mathrm{i}} e^{-r_{\mathrm{i}} t} + c_{1,\mathrm{e}} e^{-r_{\mathrm{e}} t},$$

with $c_{1,\mathrm{i}}$ and $c_{1,\mathrm{e}}$ holding costs corresponding to each type of inflation, and the remaining notation is as in section 1. The alternative form

$$\mathcal{C} = \sum_{j=1}^{n} \left\{ K e^{-r_{\mathrm{i}} t_{j-1}} + \int_{t_{j-1}}^{t_j} c_1(t) I(t)\, dt + \int_{t_{j-1}}^{t_j} c_2 e^{-r_{\mathrm{e}} t} \theta(t) I(t)\, dt \right\}, \tag{5.2}$$

where the deterioration cost at any time is related to the value of the stock at that time, may also be adopted. The equivalent form of (1.16) for the OHP model with inflation and without shortages is

$$\mathcal{C} = \sum_{j=1}^{n} \left\{ K e^{-r_{\mathrm{i}} t_{j-1}} + \int_{t_{j-1}}^{t_j} c_1(t) I(t)\, dt + c_2 e^{-r_{\mathrm{e}} t_{j-1}} I(t_{j-1}) \right\}. \tag{5.3}$$

If $r_{\mathrm{i}} = 0$, then the cost functions given by (5.1)–(5.3) are amenable to the analysis developed in section 2.

For the models with shortage, the adaptation to account for inflation is to replace (4.4) in the OHD model by

$$\int_{t_{j-1}}^{t_j} c_3(t) \left| I(t) \right| \, dt,$$

where

$$c_3(t) = c_{3,\mathrm{i}} e^{-r_{\mathrm{i}} t} + c_{3,\mathrm{e}} e^{-r_{\mathrm{e}} t},$$

and $c_{3,\mathrm{i}}$ and $c_{3,\mathrm{e}}$ are shortage costs associated with internal inflation and external inflation, respectively. In the OHP model with inflation, (4.5) is generalized to

$$c_2 e^{-r_{\mathrm{e}} t_j} \left| I(t_j) \right| + \int_{t_{j-1}}^{t_j} c_3(t) \left| I(t) \right| \, dt.$$

Again, if $r_{\mathrm{i}} = 0$, then the theory of section 2 applies, and results comparable to those in the previous section are obtainable.

REFERENCES

[1] R. C. Baker and T. L. Urban, *A deterministic inventory system with an inventory-level-dependent demand rate*, J. Oper. Res. Soc., 39 (1988), pp. 823–831.

[2] Z. T. Balkhi and L. Benkherouf, *On an inventory model for deteriorating items with stock dependent and time-varying demand rates*, Comput. Oper. Res., 31 (2004), pp. 223–240.

[3] L. Benkherouf, *On an inventory model with deteriorating items and decreasing time-varying demand and shortages*, European J. Oper. Res., 86 (1995), pp. 293–299.

[4] L. Benkherouf, *On the optimality of a replenishment policy for an inventory model with deteriorating items and time-varying demand and shortages*, Arab J. Math. Sci., 3 (1997), pp. 59–67.

[5] L. Benkherouf and Z. T. Balkhi, *On an inventory model for deteriorating items and time-varying demand*, Math. Methods Oper. Res., 45 (1997), pp. 221–233.

[6] L. Benkherouf and M. G. Mahmoud, *On an inventory model for deteriorating items with increasing time-varying demand and shortages*, J. Oper. Res. Soc., 47 (1996), pp. 188–200.

[7] M. Corstjens and P. Doyle, *A model for optimizing retail space allocations*, Manag. Sci., 27 (1981), pp. 822–833.

[8] T. K. Datta and A. K. Pal, *A note on an inventory model with inventory–level-dependent demand rate*, J. Oper. Res. Soc., 41 (1990), pp. 971–975.

[9] W. A. Donaldson, *Inventory replenishment policy for a linear trend in demand—An analytical solution*, Oper. Res. Quart., 28 (1977), pp. 663–670.

[10] P. M. Ghare and G. F. Schrader, *A model for an exponentially decaying inventory*, J. Indust. Eng., 14 (1963), pp. 238–243.

[11] S. K. Goyal and B. C. Giri, *Recent trends in modeling of deteriorating inventory*, European J. Oper. Res., 134 (2001), pp. 1–16.

[12] R. J. Henery, *Inventory replenishment policy for increasing demand*, J. Oper. Res. Soc., 30 (1979), pp. 611–617.

[13] S. Pal, A. Goswami, and K. S. Chaudhuri, *A deterministic inventory model for deteriorating items with stock-dependent demand rate*, Int. J. Prod. Econ., 32 (1993), pp. 291–299.

[14] S. Papachristos and K. Skouri, *An optimal replenishment policy for deteriorating items with time-varying demand and partial–exponential type–backlogging*, Oper. Res. Lett., 27 (2000), pp. 175–184.

[15] F. Raafat, *Survey of literature on continuously deteriorating inventory models*, J. Oper. Res. Soc., 42 (1991), pp. 27–37.

[16] P. B. Schary and B. W. Becker, *Distribution and final demand: The influence of availability*, Mississippi Valley J. Bus. Econ., 8 (1972), pp. 17–26.

[17] J. T. Teng, M. S. Chern, H. L. Yang, and Y. J. Wang, *Deterministic lot-size inventory models with shortages and deterioration for fluctuating demand*, Oper. Res. Lett., 24 (1999), pp. 65–72.

[18] H. L. Yang, J. T. Teng, and M. S. Chern *Deterministic inventory lot-size models under inflation with shortages and deterioration for fluctuating demand*, Naval Res. Logist., 48 (2001), pp. 144–158.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 2, pp. 1031–1054

© 2009 Society for Industrial and Applied Mathematics

INPUT-TO-STATE STABILITY OF DIFFERENTIAL INCLUSIONS WITH APPLICATIONS TO HYSTERETIC AND QUANTIZED FEEDBACK SYSTEMS[*]

BAYU JAYAWARDHANA[†], HARTMUT LOGEMANN[‡], AND EUGENE P. RYAN[‡]

Abstract. Input-to-state stability (ISS) of a class of differential inclusions is proved. Every system in the class is of Lur'e type: a feedback interconnection of a linear system and a set-valued nonlinearity. Applications of the ISS results, in the context of feedback interconnections with a hysteresis operator or a quantization operator in the feedback path, are developed.

Key words. absolute stability, circle criterion, differential inclusions, input-to-state stability, hysteresis, nonlinear systems, quantization

AMS subject classifications. 34A60, 34C55, 34D05, 47J40, 93C10, 93D05, 93D10

DOI. 10.1137/070711323

1. Introduction.

Classical absolute stability theory, with origins in [18], is concerned with the analysis of systems of Lur'e type, that is, feedback interconnections of the form shown in Figure 1.1, consisting of a linear system L in the forward path and a static sector-bounded nonlinearity f in the (negative) feedback path. The methodology seeks to conclude stability of the overall system through the interplay or reciprocation of inherent frequency-domain properties of the linear component L and sector data for the nonlinearity f. Accounts of the classical theory can be found in, e.g., [7, 10, 13, 19, 21, 23]. The present paper adopts a similar standpoint but differs from the classical framework in three fundamental aspects: (i) in contrast with the literature, wherein the focus is on global asymptotic stability and L^2 or L^∞ stability, input-to-state stability (ISS) issues are addressed here; (ii) nonlinearities of considerably greater generality are permitted in the feedback path; (iii) the sector conditions of the classical theory are significantly weakened. With reference to (i), conditions on the linear and nonlinear components are identified under which ISS of the interconnection is guaranteed. With reference to (ii), a framework is developed of sufficient generality to encompass not only static nonlinearities but also causal operators (and hysteresis, in particular) and quantization operators in the feedback path. With reference to (iii), through the concept of a generalized sector condition, the investigation is extended to include nonlinearities which satisfy a sector condition only in the complement of a compact set: a theory is developed pertaining to ISS *with bias*. We proceed to outline these features more precisely.

With reference to Figure 1.2, the focus of the paper is a study of absolute stability, ISS, and boundedness properties of a feedback interconnection of a finite-dimensional, linear, m-input, m-output system (A, B, C) and a set-valued nonlinearity Φ. Throughout, we assume that Δ is a set-valued map in which input or disturbance signals are

[*] Received by the editors December 19, 2007; accepted for publication (in revised form) October 7, 2008; published electronically March 6, 2009. Supported by the UK Engineering & Physical Sciences Research Council grant GR/S94582/01.

http://www.siam.org/journals/sicon/48-2/71132.html

[†] Department of Discrete Technology and Production Automation, University of Groningen, 9747 AG Groningen, The Netherlands (B.Jayawardhana@rug.nl).

[‡] Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK (hl@maths.bath.ac.uk, epr@maths.bath.ac.uk).

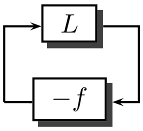


Fig. 1.1. *Classical feedback interconnection.*

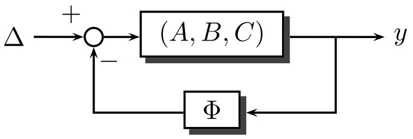


Fig. 1.2. *Interconnection of a linear system* (A, B, C) *and a set-valued nonlinearity* Φ.

embedded. We seek an analytical framework of sufficient generality to encompass inter alia feedback systems with causal operators (and, in particular, hysteresis operators) in the feedback loop. To illustrate this, let F be a causal operator from $\mathrm{dom}(F) \subset L^1_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R}^m)$ to $L^1_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R}^m)$, where $\mathbb{R}_+ := [0, \infty)$ and consider the feedback system (structurally of Lur'e type), with input $d \in L^\infty_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R}^m)$, given by the functional differential equation

$$\dot{x}(t) = Ax(t) + B\big(d(t) - (F(Cx))(t)\big). \tag{1.1}$$

By causality of F, we mean that, for all $y, z \in \mathrm{dom}(F)$ and all $\alpha > 0$,

$$y|_{[0,\alpha]} = z|_{[0,\alpha]} \implies F(y)|_{[0,\alpha]} = F(z)|_{[0,\alpha]}.$$

To associate (1.1) with the structure of Figure 1.2, assume that F can be embedded in a set-valued map Φ in the sense that

$$y \in \mathrm{dom}(F) \quad \implies \quad (F(y))(t) \in \Phi(y(t)) \;\text{ for a.e. } t \in \mathbb{R}_+.$$

If the input d is such that $d(t) \in \Delta(t)$ for almost all t, then any solution of (1.1) is a fortiori a solution of the feedback interconnection in Figure 1.2. In this sense, properties of solutions of the feedback interconnection are inherited by solutions of (1.1). Under particular regularity assumptions on Δ and Φ, generalized sector conditions on Φ, and positive-real conditions related to the linear component (A, B, C), we establish ISS (in the sense of [20] but extended to differential inclusions) and boundedness properties of solutions of the system in Figure 1.2. The approach is partially based on that of Arcak & Teel [1]. In particular, some of the arguments adopted in the proof of Lemma 5.1 of the present paper are generalizations, to a differential inclusions setting, of arguments in [1]. The paper is structured as follows. In section 2, we make precise the nature of the maps Φ and Δ and state an existence theorem which underpins the stability analysis of the differential inclusion formulation implicit in Figure 1.2. The main results, Theorems 3.4 and 3.5 (and Corollaries 3.6 and 3.7), are assembled in section 3. For clarity of presentation, the proof of Theorem 3.4 (respectively, Theorem 3.5) is presented separately in section 4 (respectively, section 5). In section 6, the results in Theorem 3.4/Corollary 3.6 are applied in the context of single-input, single-output feedback interconnections with a hysteresis operator F in the feedback loop. New absolute stability and boundedness results are obtained for Lur'e systems with Preisach hysteresis (see, e.g., [3, 9, 12, 16, 17] for earlier stability results for hysteretic feedback systems). In the final section, quantized feedback systems are considered: these constitute an area of growing importance (see, e.g., [4, 8] in a linear systems context). Specifically, in section 7, nonlinear feedback systems with uniform output quantization (parameterized by $\gamma \geq 0$) are investigated. Through an application of Theorem 3.5/Corollary 3.7, we establish robustness with respect to quantization in the following sense: if, in the absence of quantization ($\gamma = 0$), the feedback system is ISS, then, in the presence of quantization ($\gamma > 0$), the feedback system is ISS *with bias* and is such that the unbiased ISS property of the unquantized system is "approached" as $\gamma \downarrow 0$.

Notation and terminology. The open right-half complex plane is denoted by $\mathbb{C}_+$. For nonempty $S \subset \mathbb{R}^m$, we define $|S| := \sup\{\|s\| \mid s \in S\}$. If H is a proper real-rational matrix of format $m \times m$, then we say that H is *positive real* if

$$H(s) + H^*(s) \geq 0\,, \quad \forall\, s \in \mathbb{C}_+,\ s \text{ not a pole of } H,$$

where $H^*(s) := (H(s))^*$. Moreover, if $H \in H^\infty := H^\infty(\mathbb{C}_+, \mathbb{C}^{m\times m})$ (and so H does not have any poles in $\overline{\mathbb{C}}_+$), then

$$\|H\|_{H^\infty} := \sup_{s\in\mathbb{C}_+} \|H(s)\|\,,$$

where $\|H(s)\|$ is the matrix norm induced by the 2-norm on $\mathbb{C}^m$. Let $\mathcal{K}$ denote the set of all continuous and strictly increasing functions $f : \mathbb{R}_+ \to \mathbb{R}_+$, with $f(0) = 0$. We say that a function f is in $\mathcal{K}_\infty$ if $f \in \mathcal{K}$ and $f(s) \to \infty$ as $s \to \infty$. Finally, $\mathcal{KL}$ denotes the class of all continuous functions $f : \mathbb{R}_+^2 \to \mathbb{R}_+$ such that, for each $r \in \mathbb{R}_+$, the function $s \mapsto f(r, s)$ is in $\mathcal{K}$ and, for each $s \in \mathbb{R}_+$, the function $r \mapsto f(r, s)$ is nonincreasing with $f(r, s) \to 0$ as $r \to \infty$.

2. Set-valued nonlinearities and differential inclusions.

A set-valued map $y \mapsto \Phi(y) \subset \mathbb{R}^m$, with nonempty values and defined on $\mathbb{R}^m$, is said to be *upper semicontinuous at* $y \in \mathbb{R}^m$ if, for every open set U containing $\Phi(y)$, there exists an open neighborhood Y of y such that $\Phi(Y) := \cup_{z\in Y}\Phi(z) \subset U$; the map Φ is said to be *upper semicontinuous* if it is upper semicontinuous at every $y \in \mathbb{R}^m$. The set of upper semicontinuous compact–convex-valued maps

$$\Phi : \mathbb{R}^m \to \{S \subset \mathbb{R}^m \mid S \text{ nonempty, compact, and convex}\}$$

is denoted by $\mathcal{U}$. Let $\Delta : \mathbb{R}_+ \to \{S \subset \mathbb{R}^m \,|\, S \neq \emptyset\}$ be a set-valued map. The map Δ is said to be *measurable* if the preimage $\Delta^{-1}(U) := \{t \in \mathbb{R}_+ \,|\, \Delta(t) \cap U \neq \emptyset\}$ of every open set $U \subset \mathbb{R}^m$ is Lebesgue measurable; Δ is said to be *locally essentially bounded* if Δ is measurable and the function $t \mapsto |\Delta(t)|$ is in $L^\infty_{\text{loc}}(\mathbb{R}_+)$. The set of all locally essentially bounded set-valued maps $\mathbb{R}_+ \to \{S \subset \mathbb{R}^m \,|\, S \neq \emptyset\}$ is denoted by $\mathcal{B}$. For $\Delta \in \mathcal{B}$, $I \subset \mathbb{R}_+$ an interval, and $1 \leq p \leq \infty$, the L^p-norm of the restriction of the function $t \mapsto |\Delta(t)|$ to the interval I is denoted by $\|\Delta\|_{L^p(I)}$. For later use, we record a technicality.

LEMMA 2.1. *Assume that* $\Phi \in \mathcal{U}$, $\Phi(0) = \{0\}$, *and there exists* $\varphi \in \mathcal{K}_\infty$, *with*

$$\varphi(\|y\|)\|y\| \leq \langle y, v\rangle \quad \forall\, v \in \Phi(y)\ \forall\ y \in \mathbb{R}^m.$$

Then there exists $\psi \in \mathcal{K}_\infty$ *such that*

$$\|v\| \leq \psi(\|y\|) \quad \forall\, v \in \Phi(y)\ \forall\ y \in \mathbb{R}^m.$$

Proof. By upper semicontinuity of Φ and compactness of its values, for every compact set $K \subset \mathbb{R}^m$, the set $\Phi(K)$ is compact (see, for example, [2, Proposition 3,

p. 42]), and so the function $s \mapsto \psi_0(s) := \max\{\|v\| \,|\, v \in \Phi(y),\ \|y\| \le s\}$ is well defined and nondecreasing on $\mathbb{R}_+$, with $\psi_0(0) = 0$. Clearly, $\varphi(s) \le \psi_0(s)\ \forall\ s \in \mathbb{R}_+$, and so $\psi_0(s) \to \infty$ as $s \to \infty$. Let $\psi \in \mathcal{K}_\infty$ be such that $\psi(s) \ge \psi_0(s)\ \forall\ s \in \mathbb{R}_+$, for example, the function $\psi \in \mathcal{K}_\infty$ given by

$$\psi(0) := 0, \quad \psi(s) := \frac{1}{s}\int_s^{2s} \psi_0(\sigma)\mathrm{d}\sigma \quad \forall\ s > 0$$

suffices. □

The feedback system shown in Figure 1.2 corresponds to the initial-value problem

$$\dot{x}(t) - Ax(t) \in B\left(\Delta(t) - \Phi(Cx(t))\right), \quad x(0) = x^0 \in \mathbb{R}^n, \ \ \Delta \in \mathcal{B}, \tag{2.1}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{m\times n}$, and $\Phi \in \mathcal{U}$. By a solution of (2.1), we mean an absolutely continuous function $x : [0,\omega) \to \mathbb{R}^n$, $0 < \omega \le \infty$, such that $x(0) = x^0$ and the differential inclusion in (2.1) is satisfied almost everywhere on $[0,\omega)$; a solution is *maximal* if it has no proper right extension that is also a solution; a solution is *global* if it exists on $[0,\infty)$. Before developing a stability theory for systems of the form (2.1), we briefly digress to record an existence result.

LEMMA 2.2. *Let $\Phi \in \mathcal{U}$. For each $x^0 \in \mathbb{R}^n$ and each $\Delta \in \mathcal{B}$, initial-value problem* (2.1) *has a solution. Moreover, every solution can be extended to a maximal solution $x : [0,\omega) \to \mathbb{R}^n$, and if x is bounded, then x is global.*

Proof. Let $x^0 \in \mathbb{R}^n$ and $\Delta \in \mathcal{B}$ be arbitrary. By [6, Corollary 5.2], initial-value problem (2.1) has a solution, and every solution can be extended to a solution $x : [0,\omega) \to \mathbb{R}^n$ with the property that the graph of x is unbounded. Evidently, x is maximal, and if x is bounded, then $\omega = \infty$. □

3. ISS: The main results.

In the context of differential inclusion (2.1), the transfer-function matrix of the linear system given by (A, B, C) is denoted by G, i.e., $G(s) = C(sI - A)^{-1}B$.

We assemble four hypotheses which will be variously invoked in the theory developed below.

(H1) There exist numbers $a < b$ and $\delta > 0$ such that

$$\langle ay - v, by - v\rangle \le 0 \quad \forall\, v \in \Phi(y)\ \forall\, y \in \mathbb{R}^m, \tag{3.1}$$

$G(I + aG)^{-1} \in H^\infty$, and $(I + bG)(I + aG)^{-1} - \delta I$ is positive real.

(H2) $\Phi(0) = \{0\}$ and there exist numbers $a > 0$, $\delta \in [0, 1)$, and $\theta \ge 0$ such that

$$a\|y\|^2 \le \langle y, v\rangle \quad \forall\, v \in \Phi(y)\ \forall\, y \in \mathbb{R}^m, \tag{3.2}$$

$$\|v - a\delta y\| \le \langle y, v - a\delta y\rangle \quad \forall\, v \in \Phi(y)\ \forall\, y \in \mathbb{R}^m, \text{ with } \|y\| \ge \theta, \tag{3.3}$$

and $G(I + \delta aG)^{-1}$ is positive real.

(H3) There exist $\varphi \in \mathcal{K}_\infty$ and numbers $b > 0$ and $\delta \in [0, 1)$ such that

$$\max\left\{\varphi(\|y\|)\|y\|, \|v\|^2/b\right\} \le \langle y, v\rangle \quad \forall\, v \in \Phi(y)\ \forall\, y \in \mathbb{R}^m, \tag{3.4}$$

and $(\delta/b)I + G$ is positive real.

(H4) $\Phi(0) = \{0\}$ and there exist $\varphi \in \mathcal{K}_\infty$ and a number $\theta \ge 0$ such that

$$\varphi(\|y\|)\|y\| \le \langle y, v\rangle \quad \forall\, v \in \Phi(y)\ \forall\, y \in \mathbb{R}^m, \tag{3.5}$$

$$\|v\| \le \langle y, v\rangle \quad \forall\, v \in \Phi(y)\ \forall\, y \in \mathbb{R}^m, \text{ with } \|y\| \ge \theta, \tag{3.6}$$

and G is positive real.

Remark 3.1. (a) (H1) is a set-valued version of the familiar multivariable sector condition. A routine calculation shows that (3.1) holds if and only if

$$\left\| v - \frac{a+b}{2} y \right\| \le \frac{b-a}{2} \|y\| \quad \forall\, v \in \Phi(y)\ \forall\, y \in \mathbb{R}^m.$$

(b) If $m = 1$ (the single-input, single-output case), then the combined frequency-domain assumptions in (H1), namely, the condition $G(I+aG)^{-1} \in H^\infty$ together with the positive realness of $(I+bG)(I+aG)^{-1} - \delta I$, admit a graphical characterization in terms of the Nyquist diagram of G (see, e.g., [13, pp. 268]).
(c) Conditions (3.2) and (3.5) can be viewed as the limits of (3.1) and (3.4), respectively, as $b \to \infty$.
(d) A sufficient condition for (3.4) to hold is the "nonlinear" sector condition

$$\big\langle \varphi(y)\|y\|^{-1} y - v\,,\, by - v \big\rangle \le 0 \quad \forall\, v \in \Phi(y)\ \forall\, y \in \mathbb{R}^m, \tag{3.7}$$

which is (3.1) with the term ay replaced by $\varphi(y)\|y\|^{-1}y$ (which should be interpreted as taking the value 0 for $y = 0$). It is easy to construct counterexamples which show that (3.7) is not necessary for (3.4) to hold.
(e) If $m = 1$ and (3.2) holds, then (3.3) is trivially satisfied for any $\theta \ge 1$ and any $\delta \in [0, 1)$. Similarly, if $m = 1$ and (3.5) holds, then (3.6) is satisfied for every $\theta \ge 1$.
(f) If (3.4) holds for some $\varphi \in \mathcal{K}_\infty$ and for some $b > 0$, then $\Phi(0) = \{0\}$, and furthermore, (3.6) is satisfied for any $\theta > 0$ satisfying $\varphi(\theta) \ge b$.

Definition 3.2. *System* (2.1) *is said to be* input-to-state stable with bias $c \ge 0$ *if every maximal solution of* (2.1) *is global and there exist* $\beta_1 \in \mathcal{KL}$ *and* $\beta_2 \in \mathcal{K}_\infty$ *such that, for all* $x^0 \in \mathbb{R}^n$ *and all* $\Delta \in \mathcal{B}$*, every global solution* x *satisfies*

$$\|x(t)\| \le \max\big\{\beta_1(t, \|x^0\|), \beta_2(\|\Delta\|_{L^\infty[0,t]} + c)\big\} \quad \forall\, t \in \mathbb{R}_+. \tag{3.8}$$

System (2.1) *is* input-to-state stable *if it is input-to-state stable with bias* 0.

System (2.1) has the *converging-input-converging-state property* if, for all $x^0 \in \mathbb{R}^n$ and all $\Delta \in \mathcal{B}$ with $\|\Delta\|_{L^\infty[t,\infty)} \to 0$ as $t \to \infty$, every maximal solution x of (2.1) is global and satisfies $x(t) \to 0$ as $t \to \infty$. The following lemma shows in particular that if system (2.1) is input-to-state stable, then it has the converging-input-converging-state property.

Lemma 3.3. *Assume that system* (2.1) *is input-to-state stable with bias* $c \ge 0$*, and let* β_1 *and* β_2 *be as in Definition* 3.2*. Let* $x^0 \in \mathbb{R}^n$ *and* $\Delta \in \mathcal{B}$*. If* Δ *is essentially bounded* ($\|\Delta\|_{L^\infty[0,\infty)} < \infty$)*, then every global solution* x *of* (2.1) *satisfies*

$$\limsup_{t\to\infty} \|x(t)\| \le \limsup_{t\to\infty} \beta_2(\|\Delta\|_{L^\infty[t,\infty)} + c).$$

Proof. Let $x^0 \in \mathbb{R}^n$, and let $\Delta \in \mathcal{B}$ be essentially bounded. Let x be a global solution of (2.1), let $\tau \ge 0$ be arbitrary, and set $x_\tau(t) := x(t+\tau)$ and $\Delta_\tau(t) := \Delta(t+\tau)$ $\forall\, t \ge 0$. Then, $\Delta_\tau \in \mathcal{B}$ and x_τ satisfies the initial-value problem

$$\dot{x}_\tau(t) - Ax_\tau(t) \in B(\Delta_\tau(t) - \Phi(Cx_\tau(t))), \quad x_\tau(0) = x(\tau).$$

By ISS with bias c,

$$\begin{aligned} \|x(t+\tau)\| = \|x_\tau(t)\| &\le \max\big\{\beta_1(t, \|x(\tau)\|), \beta_2(\|\Delta_\tau\|_{L^\infty[0,t]} + c)\big\} \\ &= \max\big\{\beta_1(t, \|x(\tau)\|), \beta_2(\|\Delta\|_{L^\infty[\tau,t+\tau]} + c)\big\} \quad \forall\, t \in \mathbb{R}_+. \end{aligned}$$

Therefore, $\limsup_{t\to\infty}\|x(t)\| \le \beta_2\big(\|\Delta\|_{L^\infty[\tau,\infty)} + c\big)\ \forall\ \tau \ge 0$, from which the claim follows. □

We now state the two main results on ISS. The proofs can be found in sections 4 and 5.

THEOREM 3.4. *Let linear system (A, B, C) be stabilizable and detectable. Assume that* (H1) *holds. Then, every maximal solution of* (2.1) *is global and there exist positive constants c_1, c_2, and ε such that, for all $x^0 \in \mathbb{R}^n$ and $\Delta \in \mathcal{B}$, every global solution x satisfies*

$$\|x(t)\| \le c_1 e^{-\varepsilon t}\|x^0\| + c_2\|\Delta\|_{L^\infty[0,t]} \quad \forall\, t \in \mathbb{R}_+.$$

In particular, system (2.1) *is input-to-state stable.*

THEOREM 3.5. *Let linear system (A, B, C) be minimal. Assume that at least one of hypotheses* (H2), (H3), *or* (H4) *holds. Then system* (2.1) *is input-to-state stable.*

In [1] it is has been proved, for single-valued Φ and Δ, that if (H4) holds, then (2.1) is input-to-state stable. Therefore, Theorem 3.5 can be considered as a generalization of the main result in [1].

In the following two corollaries (to Theorem 3.4 and Theorem 3.5, respectively), we will consider not only nonlinearities satisfying at least one of the conditions (3.1), (3.2), (3.4), and (3.5) *for all* arguments $y \in \mathbb{R}^m$, but also nonlinearities $\Phi \in \mathcal{U}$ with the property that there exist a set-valued map $\tilde{\Phi} \in \mathcal{U}$ satisfying at least one of the conditions (3.1), (3.2), (3.4), and (3.5) and a compact set $K \subset \mathbb{R}^m$ such that

$$y \in \mathbb{R}^m\backslash K \quad\Longrightarrow\quad \Phi(y) \subset \tilde{\Phi}(y). \tag{3.9}$$

For example, single-input, single-output hysteretic elements can be subsumed by this set-valued formulation provided that the "characteristic diagram" of the hysteresis is contained in the graph of some $\Phi \in \mathcal{U}$; see section 6 for details.

COROLLARY 3.6. *Let linear system (A, B, C) be stabilizable and detectable. Let $\Phi \in \mathcal{U}$ be such that there exist a set-valued map $\tilde{\Phi} \in \mathcal{U}$ and a compact set $K \subset \mathbb{R}^m$ such that* (3.9) *holds. Assume that* (H1) *holds with Φ replaced by $\tilde{\Phi}$. Then, every maximal solution of* (2.1) *is global and there exist positive constants c_1, c_2, and ε such that, for all $x^0 \in \mathbb{R}^n$ and $\Delta \in \mathcal{B}$, every global solution x satisfies*

$$\|x(t)\| \le c_1 e^{-\varepsilon t}\|x^0\| + c_2(\|\Delta\|_{L^\infty[0,t]} + E) \quad \forall\, t \in \mathbb{R}_+,$$

where

$$E := \sup\nolimits_{y\in K} \sup\nolimits_{v\in\Phi(y)} \inf\nolimits_{\tilde{v}\in\tilde{\Phi}(y)} \|v - \tilde{v}\|. \tag{3.10}$$

Proof. First, we remark that, by upper semicontinuity of Φ and $\tilde{\Phi} \in \mathcal{U}$, together with compactness of their values and compactness of K, E is finite. Let $x^0 \in \mathbb{R}^n$ and $\Delta \in \mathcal{B}$. By Lemma 2.2, (2.1) has a solution, and every solution can be maximally extended. Let $x : [0, \omega) \to \mathbb{R}^n$ be a maximal solution of (2.1) and write $y := Cx$. Define $z \in L^1_{\mathrm{loc}}([0, \omega), \mathbb{R}^n)$ by $z := \dot{x} - Ax$. Since $z(t) \in B\big(\Delta(t) - \Phi(Cx(t))\big)$ for almost every $t \in [0, \omega)$, there exist functions $d, v : [0, \omega) \to \mathbb{R}^m$ such that

$$(d(t), v(t)) \in \Delta(t) \times \Phi(y(t)) \ \ \forall\ t \in [0, \omega), \quad z(t) = B\big(d(t) - v(t)\big) \ \text{ for a.e. } t \in [0, \omega).$$

For each $t \in [0, \omega)$, let $\tilde{v}(t) \in \tilde{\Phi}(y(t))$ be the unique point of the closed convex set $\tilde{\Phi}(y(t))$ closest to $v(t) \in \Phi(y(t))$. Then

$$y(t) \in K \implies \|v(t) - \tilde{v}(t)\| \le E, \qquad y(t) \in \mathbb{R}^m\backslash K \implies \|v(t) - \tilde{v}(t)\| = 0.$$

Define $\tilde{\Delta} \in \mathcal{B}$ by $\tilde{\Delta}(t) := \Delta(t) + \mathbb{B}_E$ (where $\mathbb{B}_E$ denotes the ball of radius $E > 0$ centered at 0 in $\mathbb{R}^m$) and $\tilde{d} : [0, \omega) \to \mathbb{R}^m$ by $\tilde{d}(t) := d(t) - v(t) + \tilde{v}(t)$. Then

$$z(t) = B(\tilde{d}(t) - \tilde{v}(t)), \quad \tilde{d}(t) \in \tilde{\Delta}, \quad \tilde{v}(t) \in \tilde{\Phi}(y(t)) \text{ for a.e. } t \in [0, \omega),$$

and so the solution x of (2.1) is also a solution of

$$\dot{x}(t) - Ax(t) \in B\big(\tilde{\Delta}(t) - \tilde{\Phi}(Cx(t))\big), \quad x(0) = x^0. \tag{3.11}$$

An application of Theorem 3.4 to (3.11) yields the claim. □

Corollary 3.7. *Let linear system (A, B, C) be minimal, and let $\Phi \in \mathcal{U}$ be such that there exist a set-valued map $\tilde{\Phi} \in \mathcal{U}$ and a compact set $K \subset \mathbb{R}^m$ such that* (3.9) *holds. Assume that at least one of the hypotheses* (H2), (H3), *or* (H4) *holds with Φ replaced by $\tilde{\Phi}$. Then system* (2.1) *is input-to-state stable with bias E, where the constant E is given by* (3.10).

Proof. The proof is identical to that of Corollary 3.6 with one exception: instead of invoking Theorem 3.4 at the end of the proof, an application of Theorem 3.5 to (3.11) completes the argument here. □

Remark 3.8. If the hypotheses of Corollary 3.6 (respectively, Corollary 3.7) hold, then there exist positive constants c_1, c_2, ε (respectively, functions $\beta_1 \in \mathcal{KL}$ and $\beta_2 \in \mathcal{K}_\infty$) such that (3.8) holds, with $c = E$ given by (3.10). We emphasize that c_1, c_2, ε (respectively, β_1 and β_2) are determined by data associated with only (A, B, C) and $\tilde{\Phi}$. In particular, they do not depend on Φ. This observation is of importance in the analysis of quantized feedback systems in section 7.

4. Proof of Theorem 3.4.

The following lemma will play an essential role in the proof of Theorem 3.4

Lemma 4.1. *Let $a < b$ and set $\kappa := (a+b)/2$ and $\lambda := (b-a)/2$. If $G(I+aG)^{-1} \in H^\infty$ and there exists $\delta > 0$ such that $(I + bG)(I + aG)^{-1} - \delta I$ is positive real, then $G(I + \kappa G)^{-1} \in H^\infty$ and $\|G(I + \kappa G)^{-1}\|_{H^\infty} < 1/\lambda$.*

Proof. Setting $\eta := \|G(I + aG)^{-1}\|_{H^\infty}$, we have that

$$(I + aG^*(s))^{-1}G^*(s)G(I + aG(s))^{-1} \le \eta^2 I \quad \forall\, s \in \mathbb{C}_+.$$

By hypothesis,

$$(I + bG(s))(I + aG(s))^{-1} + (I + aG^*(s))^{-1}(I + bG^*(s)) \ge 2\delta I \quad \forall s \in \mathbb{C}_+.$$

Setting $\varepsilon := \delta/\eta^2$, we obtain that

$$\begin{aligned} &2\varepsilon(I + aG^*(s))^{-1}G^*(s)G(s)(I + aG(s))^{-1} \\ &\qquad \le (I + bG(s))(I + aG(s))^{-1} + (I + aG^*(s))^{-1}(I + bG^*(s)) \quad \forall s \in \mathbb{C}_+. \end{aligned}$$

Therefore,

$$2\varepsilon G^*(s)G(s) \le 2I + (a + b)G^*(s) + (a + b)G(s) + 2abG^*(s)G(s) \quad \forall s \in \Gamma,$$

where $\Gamma := \{s \in \mathbb{C}_+ \mid s \text{ not a pole of } G\}$. Consequently,

$$-(ab - \varepsilon)G^*(s)G(s) \le I + \kappa G^*(s) + \kappa G(s) \quad \forall s \in \Gamma.$$

Setting $\rho := \sqrt{1 + \varepsilon/\lambda^2}$, it follows that

$$\begin{aligned} \lambda^2\rho^2 G^*(s)G(s) &\le I + \kappa G^*(s) + \kappa G(s) + \kappa^2 G^*(s)G(s) \\ &= (I + \kappa G^*(s))(I + \kappa G(s)) \quad \forall s \in \Gamma, \end{aligned}$$

which, in turn, implies that

$$\rho^2(I+\kappa G^*(s))^{-1}G^*(s)G(s)(I+\kappa G(s))^{-1} \le \lambda^{-2}I \quad \forall s \in \Gamma_0,$$

where $\Gamma_0 := \{s \in \Gamma \mid \det(sI+\kappa G(s)) \neq 0\}$. We may now infer that $G(I+\kappa G)^{-1} \in H^\infty$, and since $\rho > 1$, $\|G(I+\kappa G)^{-1}\|_{H^\infty} < 1/\lambda$. □

Proof of Theorem 3.4. Let x be a maximal solution of (2.1) defined on the maximal interval of existence $[0,\omega)$, where $0 < \omega \le \infty$. We first show that $\omega = \infty$. Seeking a contradiction, suppose that $\omega < \infty$. A routine application of the generalized Filippov selection theorem (see [22], p. 72) shows that there exists a measurable function $w : [0,\omega) \to \mathbb{R}^m$ such that $w(t) \in \Delta(t) - \Phi(Cx(t))$ for a.e. $t \in [0,\omega)$ and

$$\dot{x}(t) = Ax(t) + Bw(t) \quad \text{a.e. } t \in [0,\omega).$$

Setting $\kappa := (a+b)/2$ and $A_\kappa := A - \kappa BC$, we have

$$x(t) = e^{A_\kappa t}x^0 + \int_0^t e^{A_\kappa(t-\tau)}B(w(\tau)+\kappa Cx(\tau))\mathrm{d}\tau \quad \forall\, t \in [0,\omega). \tag{4.1}$$

Since $w(t) \in \Delta(t) - \Phi(Cx(t))$ for a.e. $t \in [0,\omega)$, there exist functions $d, v : [0,\omega) \to \mathbb{R}^m$ (not necessarily measurable) such that $w(t) = d(t) - v(t)$, $d(t) \in \Delta(t)$ and $v(t) \in \Phi(Cx(t))$ for a.e. $t \in [0,\omega)$. Setting $\lambda := (b-a)/2$ and invoking the sector condition (3.1) combined with part (a) of Remark 3.1, we may infer that

$$\begin{aligned} \|w(\tau)+\kappa Cx(\tau)\| &= \|d(\tau) - (v(\tau) - \kappa Cx(\tau))\| \\ &\le \|d(\tau)\| + \|(v(\tau)-\kappa Cx(\tau))\| \le |\Delta(\tau)| + \lambda\|Cx(\tau)\| \text{ for a.e. } \tau \in [0,\omega). \end{aligned} \tag{4.2}$$

Therefore,

$$\|x(t)\| \le \|e^{A_\kappa t}x^0\| + \|B\|\int_0^t \|e^{A_\kappa(t-\tau)}\||\Delta(\tau)|\mathrm{d}\tau + \lambda\|B\|\|C\|\int_0^t \|e^{A_\kappa(t-\tau)}\|\|x(\tau)\|\mathrm{d}\tau \quad \forall\, t \in [0,\omega).$$

Since (by supposition) ω is finite, we conclude that, for some constant $c > 0$,

$$\|x(t)\| \le c\left(1 + \int_0^t \|x(\tau)\|\mathrm{d}\tau\right) \quad \forall\, t \in [0,\omega).$$

By Gronwall's lemma, it follows that the maximal solution x is bounded on $[0,\omega)$, contradicting (via Lemma 2.2) the supposition that $\omega < \infty$. Consequently, $\omega = \infty$.

Defining $G_\kappa(s) := G(I+\kappa G(s))^{-1} = C(sI - A_\kappa)^{-1}B$, it follows from (H1), via Lemma 4.1, that $G_\kappa \in H^\infty$ and $\|G_\kappa\|_{H^\infty} < 1/\lambda$. Moreover, by stabilizability and detectability, A_κ is Hurwitz. Let $\varepsilon > 0$ be sufficiently small so that $A_\kappa + \varepsilon I$ is Hurwitz and

$$\gamma := \sup_{\mathrm{Re}\, s \ge -\varepsilon} \|G_\kappa(s)\| < 1/\lambda. \tag{4.3}$$

Set $y := Cx$ and, for all $t \in \mathbb{R}_+$, define $y_\varepsilon(t) := e^{\varepsilon t}y(t)$ and $w_\varepsilon(t) := e^{\varepsilon t}w(t)$. It follows from (4.1) that

$$y_\varepsilon(t) = Ce^{(A_\kappa+\varepsilon I)t}x^0 + \int_0^t Ce^{(A_\kappa+\varepsilon I)(t-\tau)}B(w_\varepsilon(\tau)+\kappa y_\varepsilon(\tau))\mathrm{d}\tau \quad \forall\, t \in \mathbb{R}_+.$$

Setting $k_0 := \big(\int_0^\infty \|Ce^{(A_\kappa+\varepsilon I)\tau}\|^2 \mathrm{d}\tau\big)^{1/2} < \infty$, we obtain that

$$\|y_\varepsilon\|_{L^2[0,t]} \le k_0\|x^0\| + \gamma\|w_\varepsilon + \kappa y_\varepsilon\|_{L^2[0,t]} \quad \forall\, t \in \mathbb{R}_+. \tag{4.4}$$

By (4.2),

$$\|w_\varepsilon(\tau) + \kappa y_\varepsilon(\tau))\| \le |\Delta_\varepsilon(\tau)| + \lambda\|y_\varepsilon(\tau))\| \quad \text{for a.e. } \tau \in \mathbb{R}_+, \tag{4.5}$$

where $\Delta_\varepsilon(\tau) := e^{\varepsilon\tau}\Delta(\tau)\ \forall\ \tau \in \mathbb{R}_+$. From (4.3), we see that $\gamma\lambda < 1$: setting $k_1 := 1/(1-\gamma\lambda)$ and invoking (4.4) and (4.5), we have

$$\|y_\varepsilon\|_{L^2[0,t]} \le k_1\left(k_0\|x^0\| + \gamma\|\Delta_\varepsilon\|_{L^2[0,t]}\right) \quad \forall\, t \in \mathbb{R}_+. \tag{4.6}$$

By (4.1),

$$e^{\varepsilon t}x(t) = e^{(A_\kappa+\varepsilon I)t}x^0 + \int_0^t e^{(A_\kappa+\varepsilon I)(t-\tau)}B(w_\varepsilon(\tau) + \kappa y_\varepsilon(\tau))\mathrm{d}\tau \quad \forall\, t \in \mathbb{R}_+,$$

which together with (4.5) yields

$$\|x(t)\|e^{\varepsilon t} \le k_2\|x^0\| + \|B\|\int_0^t \|e^{(A_\kappa+\varepsilon I)(t-\tau)}\|(|\Delta_\varepsilon(\tau)| + \lambda\|y_\varepsilon(\tau)\|)\mathrm{d}\tau \quad \forall\, t \in \mathbb{R}_+,$$

where $k_2 := \sup_{t\ge 0}\|e^{(A_\kappa+\varepsilon I)t}\|$. Invoking Hölder's inequality to estimate the integral on the right-hand side of the above inequality, we conclude that there exists a constant $k_3 > 0$ such that

$$\|x(t)\|e^{\varepsilon t} \le k_2\|x^0\| + k_3(\|\Delta_\varepsilon\|_{L^2[0,t]} + \lambda\|y_\varepsilon\|_{L^2[0,t]}) \quad \forall\, t \in \mathbb{R}_+.$$

Combining this with (4.6), we conclude that

$$\|x(t)\|e^{\varepsilon t} \le (k_2 + \lambda k_0 k_1 k_3)\|x^0\| + k_3(1 + \lambda\gamma k_1)\|\Delta_\varepsilon\|_{L^2[0,t]} \quad \forall\, t \in \mathbb{R}_+.$$

Noting that $\|\Delta_\varepsilon\|_{L^2[0,t]} \le (e^{\varepsilon t}/\sqrt{2\varepsilon})\|\Delta\|_{L^\infty[0,t]}\ \forall\ t \in \mathbb{R}_+$, setting $c_1 := k_2 + \lambda k_0 k_1 k_3$ and $c_2 := k_3(1+\lambda\gamma k_1)/\sqrt{2\varepsilon}$, we conclude that

$$\|x(t)\| \le c_1 e^{-\varepsilon t}\|x^0\| + c_2\|\Delta\|_{L^\infty[0,t]} \quad \forall\, t \in \mathbb{R}_+.$$

This completes the proof. □

Remark 4.2. Theorem 3.4 can be considered as a refinement of the classical circle criterion (see, for example, [7, 13, 21]). In particular, it shows that, under the standard assumptions of the circle criterion, ISS is guaranteed. The exponential weighting technique used in the proof of Theorem 3.4 is well known and has been used to prove stability results of input-output type (see [7, section V.3] and the references therein). The application of this technique in an ISS context seems to be new. In particular, whilst the standard textbook version of the circle criterion for state-space systems is usually proved using Lyapunov techniques combined with the positive-real lemma (see, for example, [13, Theorem 7.1] or [21, p. 227]), the above proof of Theorem 3.4 provides an alternative, more elementary approach. Moreover, the methodology can be extended to an infinite-dimensional setting; see [11].

5. Proof of Theorem 3.5. In this section, we provide a proof of Theorem 3.5. In contrast to the proof of Theorem 3.4, we adopt a Lyapunov argument. In particular, we prove Theorem 3.5 by establishing the existence of a Lyapunov function with special properties (a so-called ISS Lyapunov function) if any one of hypotheses (H2), (H3), or (H4) hold. This we do in two preliminary lemmas.

LEMMA 5.1. *Let linear system* (A, B, C) *be minimal. Assume that either* (H3) *or* (H4) *holds. Then there exist* $\alpha_1, \alpha_2, \alpha_3, \alpha_4 \in \mathcal{K}_\infty$, *and a continuously differentiable function* $V : \mathbb{R}^n \to \mathbb{R}_+$ *such that*

$$\left.\begin{aligned} &\alpha_1(\|x\|) \le V(x) \le \alpha_2(\|x\|) \quad \forall\, x \in \mathbb{R}^n, \\ &\max_{v\in\Phi(Cx)} \langle \nabla V(x), Ax + B(d - v)\rangle \le -\alpha_3(\|x\|) + \alpha_4(\|d\|) \\ &\hspace{16em} \forall\, (x, d) \in \mathbb{R}^n \times \mathbb{R}^m. \end{aligned}\right\} \tag{5.1}$$

Proof. By Lemma 2.1, there exists $\psi \in \mathcal{K}_\infty$ such that

$$\|v\| \le \psi(\|y\|) \quad \forall\, y \in \mathbb{R}^m \quad \forall\, v \in \Phi(y). \tag{5.2}$$

(If (H3) holds, then we may take $\psi : s \mapsto bs$ in (5.2).) Combining (5.2) with either (H3) or (H4) yields

$$\varphi(\|y\|)\|y\| \le \langle y, v\rangle \le \psi(\|y\|)\|y\| \quad \forall\, y \in \mathbb{R}^m \quad \forall\, v \in \Phi(y). \tag{5.3}$$

If (H3) holds, then $(\delta/b)I + G$ is positive real for some $\delta \in [0, 1)$; if (H4) holds, then G is positive real. Introducing the following notational convenience

$$\lambda := \begin{cases} 1/b & \text{if (H3) holds,} \\ 0 & \text{otherwise,} \end{cases}$$

both possibilities are captured by the statement that $\delta\lambda I + G$ is positive real for some $\delta \in [0, 1)$. This implies, via the positive-real lemma, the existence of a real matrix L and a symmetric, positive-definite real matrix P such that

$$PA + A^T P = -L^T L, \quad PB = C^T - \sqrt{\kappa}\, L^T, \quad \kappa := 2\delta\lambda. \tag{5.4}$$

We also record that

$$\lambda\|v\|^2 \le \langle y, v\rangle \quad \forall\, v \in \Phi(y) \quad \forall\, y \in \mathbb{R}^m. \tag{5.5}$$

Now, define $V_0 : \mathbb{R}^n \to \mathbb{R}_+,\ x \mapsto \langle x, Px\rangle$. Then, invoking (5.4),

$$\begin{aligned} \langle \nabla V_0(x), Ax + B(d - v)\rangle &= 2\langle Px, Ax\rangle + 2\langle B^T Px, (d - v)\rangle \\ &\le -\|Lx\|^2 + 2\langle Cx, (d - v)\rangle - 2\sqrt{\kappa}\langle Lx, (d - v)\rangle \\ &= -\|Lx + \sqrt{\kappa}(d - v)\|^2 + \kappa\|d - v\|^2 + 2\langle Cx, (d - v)\rangle \\ &\hspace{6em} \forall\, x \in \mathbb{R}^n, \ \ \forall\, (d, v) \in \mathbb{R}^m \times \Phi(Cx)\,, \end{aligned}$$

from which, together with (5.5), we may infer

$$\begin{aligned} \langle \nabla V_0(x), Ax + B(d - v)\rangle &\le \kappa\|d\|^2 + 2\kappa\|v\|\|d\| + \kappa\|v\|^2 + 2\|y\|\|d\| - 2\langle y, v\rangle \\ &\le 2(1 + 2\delta)\|y\|\|d\| + \kappa\|d\|^2 - 2(1 - \delta)\langle y, v\rangle \\ &\hspace{4em} \forall\, x \in \mathbb{R}^n, \ \ \forall\, (d, v) \in \mathbb{R}^m \times \Phi(y), \ y = Cx. \end{aligned} \tag{5.6}$$

Observe that, for all $y \in \mathbb{R}^m$ and all $(d,v) \in \mathbb{R}^m \times \Phi(y)$,

$$\begin{aligned}
&2(1+2\delta)\|d\| \le (1-\delta)\varphi(\|y\|) \implies \\
&\qquad\qquad 2(1+2\delta)\|d\|\|y\| \le (1-\delta)\varphi(\|y\|)\|y\| \le (1-\delta)\langle y,v\rangle, \\
&2(1+2\delta)\|d\| > (1-\delta)\varphi(\|y\|) \implies \\
&\qquad\qquad 2(1+2\delta)\|d\|\|y\| < 2(1+2\delta)\|d\|\varphi^{-1}(2(1+2\delta)\|d\|/(1-\delta))
\end{aligned}$$

and so, defining $\gamma \in \mathcal{K}_\infty$ by $\gamma(s) := 2(1+2\delta)s\,\varphi^{-1}\left(2(1+2\delta)s/(1-\delta)\right)$, we have

$$2(1+2\delta)\|d\|\|y\| \le (1-\delta)\langle y,v\rangle + \gamma(\|d\|) \quad \forall\, y \in \mathbb{R}^m \quad \forall\, (d,v) \in \mathbb{R}^m \times \Phi(y). \tag{5.7}$$

The conjunction of (5.6) and (5.7) gives

$$\begin{aligned}
\langle \nabla V_0(x), Ax + B(d-v)\rangle \le -(1-\delta)\langle y,v\rangle + \gamma(\|d\|) + \kappa\|d\|^2 \\
\forall\, x \in \mathbb{R}^n,\ \forall\, (d,v) \in \mathbb{R}^m \times \Phi(y),\ y = Cx.
\end{aligned} \tag{5.8}$$

Let $H \in \mathbb{R}^{n\times m}$ be such that $A - HC$ is Hurwitz. Let $Q = Q^T > 0$ be such that

$$Q(A-HC) + (A-HC)^T Q = -3I$$

and define $W : \mathbb{R}^n \to \mathbb{R}_+$ by $W(x) := \langle x, Qx\rangle$.

Writing $k_0 := \max\left\{2\|QB\|\,,\, 2\|QH\|\,,\, \|QB\|^2\right\}$, we have

$$\begin{aligned}
\langle \nabla W(x), Ax + B(d-v)\rangle &= 2\langle Qx, (A-HC)x + Hy + B(d-v)\rangle \\
&= -3\|x\|^2 + 2\langle H^T Qx, y\rangle + 2\langle B^T Qx, d-v\rangle \\
&\le -2\|x\|^2 + k_0\|x\|\big(\|y\| + \|v\|\big) + k_0\|d\|^2 \\
&\qquad \forall\, x \in \mathbb{R}^n, \forall\, (d,v) \in \mathbb{R}^m \times \Phi(y),\ y = Cx.
\end{aligned} \tag{5.9}$$

Since either (H3) or (H4) holds and invoking part (f) of Remark 3.1 in the former case, we may infer the existence of $\theta \ge 1/2$ such that

$$y \in \mathbb{R}^m,\ \ \|y\| \ge \theta \quad \implies \quad \langle y,v\rangle \ge \|v\| \quad \forall\, v \in \Phi(y). \tag{5.10}$$

Define $f_0 \in \mathcal{K}_\infty$ by $f_0(s) := s + \psi(s)$, the continuous, nondecreasing function $f_1 : (0,\theta] \to (0,\infty)$ by

$$f_1(s) := \min_{t \in [s,\theta]} \frac{t\varphi(t)}{(f_0(t))^2},$$

and $f_2 : \mathbb{R}_+ \to \mathbb{R}_+$ by

$$f_2(s) := \begin{cases} 0, & s = 0, \\ \min\{s, f_1(s)\}, & s \in (0,\theta], \\ f_1(\theta) + (s-\theta), & s > \theta. \end{cases}$$

Observe that

$$f_1(\theta) = \frac{\theta\varphi(\theta)}{(\theta + \psi(\theta))^2} < \frac{\theta\varphi(\theta)}{2\theta\psi(\theta)} \le \frac{\theta\varphi(\theta)}{\psi(\theta)} \le \theta\,,$$

where we have used that $\theta \ge 1/2$. It follows that $f_2(\theta) = f_1(\theta)$, and therefore, f_2 is continuous. Clearly, f_2 is unbounded, and moreover, it is readily verified that f_2

is nondecreasing. Write $f_3 := f_2 \circ f_0^{-1}$ (continuous, nondecreasing, and unbounded, with $f_3(0) = 0$) and observe (for later use) that

$$\text{(5.11)} \quad \|y\| < \theta \implies f_3(\|y\| + \psi(\|y\|))(\|y\| + \psi(\|y\|))^2 = (f_3 \circ f_0)(\|y\|)(f_0(\|y\|))^2 \\ = f_2(\|y\|)(f_0(\|y\|))^2 \leq f_1(\|y\|)(f_0(\|y\|))^2 \leq \|y\|\varphi(\|y\|).$$

Next, we introduce functions $\eta \in \mathcal{K}_\infty$ and σ (continuous, nondecreasing, and unbounded, with $\sigma(0) = 0$) given by

$$\eta : s \mapsto \frac{1}{k_0}\sqrt{\frac{s}{\|Q\|}}, \quad \sigma := f_3 \circ \eta.$$

Let $s^* > 0$ be the unique point with the property $\eta(s^*)\sigma(s^*) = 1$ and define the continuous function $\rho : \mathbb{R}_+ \to \mathbb{R}_+$ by

$$\rho(s) := \begin{cases} \sigma(s), & 0 \leq s \leq s^*, \\ 1/\eta(s), & s > s^*. \end{cases}$$

Finally, define $R \in \mathcal{K}_\infty$ by

$$R(s) := \int_0^s \rho(\tau)\, d\tau,$$

and $V_1 : \mathbb{R}^n \to \mathbb{R}_+, \quad x \mapsto R(W(x))$. Note that

$$\text{(5.12)} \quad \left.\begin{array}{ll} \text{(a)} & \rho(s) \leq \sigma(s) \leq \sigma(s^*) =: k_1 \quad \forall\, s \in \mathbb{R}_+, \\ \text{(b)} & \rho(W(x))\|x\| \leq k_0\sqrt{\|Q\|\|Q^{-1}\|} =: k_2 \quad \forall\, x \in \mathbb{R}^n, \\ \text{(c)} & \rho(W(x))\|x\|^2 \geq \|x\| \min\big\{\|x\| f_3\big(\|x\|/k_2\big)\,,\, k_0\big\} \quad \forall\, x \in \mathbb{R}^n. \end{array}\right\}$$

Invoking (5.9) and (5.12)(a), we have

$$\text{(5.13)} \\ \langle \nabla V_1(x), Ax + B(d - v)\rangle \leq -2\rho(W(x))\|x\|^2 + \rho(W(x))k_0\|x\|\big(\|y\| + \|v\|\big) + k_0 k_1\|d\|^2 \\ \forall\; x \;\in \mathbb{R}^n \;\forall\; (d, v) \in \mathbb{R}^m \times \Phi(Cx).$$

We proceed to obtain a convenient estimate of the term $\rho(W(x))k_0\|x\|\big(\|y\| + \|v\|\big)$.

Write $k_3 := \frac{1}{2}\min\{1, \varphi(\theta)\}$. By (5.3) and (5.10), we have

$$\|y\| \geq \theta \implies 2\langle y, v\rangle \geq \|v\| + \|y\|\varphi(\|y\|) \geq \|v\| + \|y\|\varphi(\theta) \geq 2k_3(\|v\| + \|y\|) \\ \forall\, v \in \Phi(y),$$

which, in conjunction with (5.12)(b), gives

$$\text{(5.14)} \quad x \in \mathbb{R}^n,\; y = Cx,\; \|y\| \geq \theta \implies \\ \rho(W(x))k_0\|x\|\big(\|y\| + \|v\|\big) \leq \frac{k_0 k_2}{k_3}\langle y, v\rangle \quad \forall\, v \in \Phi(y).$$

Invoking (5.2), (5.3), and (5.11), we have

$$\|y\| < \theta \implies f_3(\|y\| + \|v\|)(\|y\| + \|v\|)^2 \\ \leq f_3(\|y\| + \psi(\|y\|))(\|y\| + \psi(\|y\|))^2 \leq \|y\|\varphi(\|y\|) \leq \langle y, v\rangle \quad \forall\, v \in \Phi(y)$$

from which, together with the observation that

$$x \in \mathbb{R}^n,\ y = Cx,\ v \in \Phi(y),\ k_0(\|y\| + \|v\|) \geq \|x\| \quad \Longrightarrow \\ \rho(W(x)) \leq \sigma(\|Q\|\|x\|^2)) \leq \sigma(k_0^2\|Q\|(\|y\| + \|v\|)^2) = f_3(\|y\| + \|v\|),$$

we may infer

$$(5.15) \quad x \in \mathbb{R}^n,\ y = Cx,\ v \in \Phi(y),\ k_0(\|y\| + \|v\|) \geq \|x\|,\ \|y\| < \theta \quad \Longrightarrow \\ \rho(W(x))k_0\|x\|\big(\|y\| + \|v\|\big) \leq \rho(W(x))\|x\|^2 + \frac{k_0^2}{4}\rho(W(x))(\|y\| + \|v\|)^2 \\ \leq \rho(W(x))\|x\|^2 + \frac{k_0^2}{4}\langle y, v\rangle\,.$$

Clearly,

$$(5.16) \quad x \in \mathbb{R}^n,\ y = Cx,\ v \in \Phi(y),\ k_0(\|y\| + \|v\|) \leq \|x\|,\ \|y\| < \theta \quad \Longrightarrow \\ \rho(W(x))k_0\|x\|\big(\|y\| + \|v\|\big) \leq \rho(W(x))\|x\|^2.$$

Combining (5.15) and (5.16), we have

$$(5.17) \quad x \in \mathbb{R}^n,\ y = Cx,\ \|y\| < \theta \quad \Longrightarrow \\ \rho(W(x))k_0\|x\|\big(\|y\| + \|v\|\big) \leq \rho(W(x))\|x\|^2 + \frac{k_0^2}{4}\langle y, v\rangle \quad \forall\ v \in \Phi(y).$$

Writing $k_4 := \max\big\{k_0k_2/k_3\,,\ k_0^2/4\big\}$, we conclude from (5.13), (5.14), (5.17) that

$$(5.18) \quad \langle \nabla V_1(x), Ax + B(d - v)\rangle \leq -\rho(W(x))\|x\|^2 + k_4\langle y, v\rangle + k_0k_1\|d\|^2 \\ \forall\ x \in \mathbb{R}^n\ \forall\ (d, v) \in \mathbb{R}^m \times \Phi(Cx).$$

Now define $V := k_4V_0 + (1-\delta)V_1$. Then, combining (5.8) and (5.18), we arrive at

$$(5.19) \quad \langle \nabla V(x), Ax + B(d - v)\rangle \\ \leq -(1-\delta)\rho(W(x))\|x\|^2 + \big((1-\delta)k_0k_1 + \kappa k_4\big)\|d\|^2 + k_4\gamma(\|d\|) \\ \forall\ x \in \mathbb{R}^n,\ \ \forall\ (d, v) \in \mathbb{R}^m \times \Phi(y),\ y = Cx.$$

Finally, defining α_1, α_2, α_3, $\alpha_4 \in \mathcal{K}_\infty$ by

$$\alpha_1(s) := k_4\|P^{-1}\|^{-1}s^2, \quad \alpha_2 := k_4\|P\|s^2 + (1-\delta)R(\|Q\|s^2),$$

$$\alpha_3(s) := (1-\delta)s\min\{sf_3(s/k_2)\,,\ k_0\}, \quad \alpha_4(s) := \big((1-\delta)k_0k_1 + \kappa k_4\big)s^2 + k_4\gamma(s),$$

and invoking (5.12)(c), we conclude that (5.1) holds. This completes the proof. □

Lemma 5.2. *Let linear system* (A, B, C) *be minimal. Assume that* (H2) *holds. Then the assertions of Lemma* 5.1 *are valid.*

Proof. Let $a > 0$, $\delta \in [0, 1)$, and $\theta \geq 0$ be as in hypothesis (H2). Without loss of generality, we may assume $\theta \geq 1/2$. Note that linear system (A_1, B, C), with $A_1 := A - \delta aBC$, is a minimal realization of $G(I + \delta aG)^{-1}$. Therefore, hypothesis (H2) implies, via the positive-real lemma, the existence of a real matrix L and a symmetric, positive-definite real matrix P such that

$$(5.20) \qquad PA_1 + A_1^TP = -L^TL, \quad PB = C^T.$$

Invoking Lemma 2.1, there exists $\psi \in \mathcal{K}_\infty$ such that (5.3) holds with $\varphi(s) = as$. Now define $\varphi_1, \psi_1 \in \mathcal{K}_\infty$, and $y \mapsto \Phi_1(y) \subset \mathbb{R}^m$ by

$$\varphi_1(s) := \varphi(s) - \delta as = (1-\delta)as, \quad \psi_1(s) := \psi(s) - \delta as \quad \forall\, s \in \mathbb{R}_+,$$

$$\Phi_1(y) := \{v - \delta ay \mid v \in \Phi(y)\} \quad \forall\, y \in \mathbb{R}^m.$$

In view of (5.3), we have

$$(5.21)\quad (1-\delta)a\|y\|^2 = \varphi_1(\|y\|)\|y\| \le \langle y, v\rangle \le \psi_1(\|y\|)\|y\| \quad \forall\, y \in \mathbb{R}^m \quad \forall\, v \in \Phi_1(y).$$

Moreover, by hypothesis (H2),

$$(5.22)\qquad y \in \mathbb{R}^m, \quad \|y\| \ge \theta \quad \implies \quad \langle y, v\rangle \ge \|v\| \quad \forall\, v \in \Phi_1(y).$$

Recalling that $A_1 := A - \delta aBC$, we have

$$(5.23)\qquad \big\{Ax - Bv \mid v \in \Phi(Cx)\big\} = \big\{A_1x - Bv \mid v \in \Phi_1(Cx)\big\} \quad \forall\, x \in \mathbb{R}^n.$$

Now, define $V_0 : \mathbb{R}^n \to \mathbb{R}_+$, $x \mapsto \langle x, Px\rangle$. Then, invoking (5.20),

$$\begin{aligned}(5.24)\quad \langle \nabla V_0(x), A_1x + B(d-v)\rangle &= 2\langle Px, A_1x\rangle + 2\langle B^T Px, (d-v)\rangle \\ &\le -\|Lx\|^2 + 2\langle Cx, (d-v)\rangle \le 2\|y\|\|d\| - 2\langle y, v\rangle \\ &\qquad \forall\, x \in \mathbb{R}^n, \ \forall\, (d,v) \in \mathbb{R}^m \times \Phi_1(y), \ y = Cx.\end{aligned}$$

Observe that, for all $y \in \mathbb{R}^m$ and all $(d,v) \in \mathbb{R}^m \times \Phi_1(y)$,

$$2\|d\| \le \varphi_1(\|y\|) \implies 2\|d\|\|y\| \le \varphi_1(\|y\|)\|y\| \le \langle y, v\rangle,$$

$$2\|d\| > \varphi_1(\|y\|) \implies 2\|d\|\|y\| < 2\|d\|\varphi_1^{-1}(2\|d\|)$$

and so, defining $\gamma \in \mathcal{K}_\infty$ by $\gamma(s) := 2s\,\varphi_1^{-1}(2s)$, it follows from (5.24) that

$$\begin{aligned}(5.25)\quad \langle \nabla V_0(x), A_1x + B(d-v)\rangle &\le -\langle y, v\rangle + \gamma(\|d\|) \\ &\qquad \forall\, x \in \mathbb{R}^n, \ \forall\, (d,v) \in \mathbb{R}^m \times \Phi_1(y), \ y = Cx.\end{aligned}$$

The conjunction of (5.23) and (5.25) yields

$$\begin{aligned}(5.26)\quad \langle \nabla V_0(x), Ax + B(d-v)\rangle &\le -\langle y, v\rangle + \delta a\|y\|^2 + \gamma(\|d\|) \\ &\qquad \forall\, x \in \mathbb{R}^n, \ \forall\, (d,v) \in \mathbb{R}^m \times \Phi(y), \ y = Cx.\end{aligned}$$

Let $H \in \mathbb{R}^{n\times m}$ be such that $A_1 - HC$ is Hurwitz. Let $Q = Q^T > 0$ be such that

$$Q(A_1 - HC) + (A_1 - HC)^T Q = -3I$$

and define $W : \mathbb{R}^n \to \mathbb{R}_+$ by $W(x) := \langle x, Qx\rangle$. The same construction as in the proof of Lemma 5.1 (with A_1 replacing A and Φ_1 replacing Φ therein) yields a function f_3 (continuous, nondecreasing, and unbounded, with $f_3(0) = 0$), a continuous function $\rho : \mathbb{R}_+ \to \mathbb{R}_+$, with primitive $R \in \mathcal{K}_\infty$, and positive constants c_0, c_1, c_2, c_3 such that, on writing $V_1 : \mathbb{R}^n \to \mathbb{R}_+$, $x \mapsto R(W(x))$, the following counterparts of (5.12)(c) and

(5.18) hold:

$$\rho(W(x))\|x\|^2 \geq \|x\| \min\left\{c_0,\, \|x\| f_3(c_1\|x\|)\right\} \quad \forall\, x \in \mathbb{R}^n, \tag{5.27}$$

$$\langle \nabla V_1(x), A_1 x + B(d-v)\rangle \leq -\rho(W(x))\|x\|^2 + c_2\langle y, v\rangle + c_3\|d\|^2$$
$$\forall\, x \in \mathbb{R}^n, \forall\, (d,v) \in \mathbb{R}^m \times \Phi_1(y),\ y = Cx.$$

In view of (5.23), the latter yields

$$\langle \nabla V_1(x), Ax + B(d-v)\rangle \leq -\rho(W(x))\|x\|^2 + c_2\langle y, v\rangle - c_2\delta a\|y\|^2 + c_3\|d\|^2 \tag{5.28}$$
$$\forall\, x \in \mathbb{R}^n, \forall\, (d,v) \in \mathbb{R}^m \times \Phi(y),\ y = Cx.$$

Now define $V := c_2 V_0 + V_1$. Then, combining (5.26) and (5.28), we have

$$\langle \nabla V(x), Ax + B(d-v)\rangle \leq -\rho(W(x))\|x\|^2 + c_2\gamma(\|d\|) + c_3\|d\|^2 \tag{5.29}$$
$$\forall\, x \in \mathbb{R}^n\ \forall\, (d,v) \in \mathbb{R}^m \times \Phi(Cx).$$

Finally, defining α_1, α_2, α_3, $\alpha_4 \in \mathcal{K}_\infty$ by

$$\alpha_1(s) := c_2\|P^{-1}\|^{-1}s^2, \quad \alpha_2 := c_2\|P\|s^2 + R(\|Q\|s^2),$$
$$\alpha_3(s) := s\min\{c_0,\, sf_3(c_1 s)\}, \quad \alpha_4(s) := c_2\gamma(s) + c_3 s^2$$

and invoking (5.27), we may conclude that (5.1) holds. This completes the proof. □

We are now in a position to prove Theorem 3.5. The argument developed below is not new and can be found (usually in form of sketch proofs) in the literature (see [20] and the references therein). For completeness, we provide a detailed proof.

Proof of Theorem 3.5. If either (H3) or (H4) holds (respectively, if (H2) holds), then Lemma 5.1 (respectively, Lemma 5.2) ensures the existence of $\alpha_1, \alpha_2, \alpha_3, \alpha_4 \in \mathcal{K}_\infty$ and continuously differentiable V such that $\alpha_1(\|x\|) \leq V(x) \leq \alpha_2(\|x\|)\ \forall\, x \in \mathbb{R}^n$ and

$$\langle \nabla V(x), Ax + B(d-v)\rangle \leq -\alpha_3(\|x\|) + \alpha_4(\|d\|) \tag{5.30}$$
$$\forall\, x \in \mathbb{R}^n\ \ \forall\, (d,v) \in \mathbb{R}^m \times \Phi(Cx).$$

Let $x^0 \in \mathbb{R}^n$ and $\Delta \in \mathcal{B}$. By Lemma 2.2, (2.1) has a solution, and every solution can be maximally extended. Let $x : [0,\omega) \to \mathbb{R}^n$ be a maximal solution of (2.1). By (5.30), we have

$$(V \circ x)'(t) \leq \alpha_4(|\Delta(t)|) \quad \text{for a.e. } t \in [0,\omega). \tag{5.31}$$

Seeking a contradiction, suppose that $\omega < \infty$. Then, by local essential boundedness of Δ and continuity of α_4, there exists $c_0 > 0$ such that $\alpha_4(|\Delta(t)|) \leq c_0\ \forall\, t \in [0,\omega)$. Now, by the final assertion of Lemma 2.2, x is unbounded, which contradicts the fact that, by (5.31), $\alpha_1(\|x(t)\|) \leq V(x(t)) \leq V(x^0) + c_0\omega\ \forall\, t \in [0,\omega)$. Therefore, every maximal solution of (2.1) is global.

Write $\alpha_5 := \alpha_3 \circ \alpha_2^{-1} \in \mathcal{K}_\infty$ and define $\alpha_6 : \mathbb{R}_+ \to \mathbb{R}_+$ by

$$\alpha_6(s) := \frac{2}{s}\int_{s/2}^{s} \alpha_5(t)dt \ \ \forall\, s > 0, \quad \alpha_6(0) := \lim_{s\downarrow 0}\alpha_6(s) = 0.$$

Since $\alpha_5 \in \mathcal{K}_\infty$, we have $\alpha_5(s/2) \le \alpha_6(s) \le \alpha_5(s)\ \forall\ s \in \mathbb{R}_+$, and moreover, α_6 is differentiable on $(0,\infty)$, with derivative $\alpha_6'(s) \ge 0\ \forall\ s \in (0,\infty)$. Now define $\alpha_7 : \mathbb{R}_+ \to \mathbb{R}_+$ by $\alpha_7(s) := \min\{1, s\}\alpha_6(s)$. Clearly, α_7 is locally Lipschitz, $\alpha_7(0) = 0$, and $0 < \alpha_7(s) \le \alpha_5(s)\ \forall\ s > 0$. Define the locally Lipschitz function

$$Z : \mathbb{R} \to \mathbb{R}, \quad \zeta \mapsto Z(\zeta) := \begin{cases} -\alpha_7(\zeta)/2, & \zeta \ge 0, \\ \alpha_7(-\zeta)/2, & \zeta < 0, \end{cases}$$

and consider the scalar system

$$\dot{z}(t) = Z(z(t)).$$

Since $Z(0) = 0$ and $\zeta Z(\zeta) = -|\zeta|\alpha_7(|\zeta|)/2 < 0\ \forall\ \zeta \ne 0$, it follows that 0 is a globally asymptotically stable equilibrium of this system which, together with the local Lipschitz property of Z, ensures the existence of a continuous global semiflow $\beta : \mathbb{R}_+ \times \mathbb{R} \to \mathbb{R}$ (and so, for each $z^0 \in \mathbb{R}$, $z : \mathbb{R}_+ \to \mathbb{R}$, $t \mapsto \beta(t, z^0)$, is the unique global solution of the initial-value problem $\dot{z} = Z(z)$, $z(0) = z^0$; moreover, $\beta(t, z^0) \to 0$ as $t \to \infty$). Let $\beta_0 := \beta|_{\mathbb{R}_+ \times \mathbb{R}_+}$ be the restriction of β to $\mathbb{R}_+ \times \mathbb{R}_+$. Evidently, $\beta_0 \in \mathcal{KL}$. Now define $\beta_1 \in \mathcal{KL}$ by

$$\beta_1(t, s) := \alpha_1^{-1}(\beta_0(t, \alpha_2(s)))$$

and define $\beta_2 \in \mathcal{K}_\infty$ by

$$\beta_2(s) := \big(\alpha_1^{-1} \circ \alpha_2 \circ \alpha_3^{-1}\big)(2\alpha_4(s)).$$

Let x^0 and $\Delta \in \mathcal{B}$ be arbitrary, and let x be a global solution of (2.1). Let $t \in \mathbb{R}_+$ be arbitrary. By (5.30), we have

$$(V \circ x)'(\tau) \le -\alpha_3(\|x(\tau)\|) + \alpha_4(|\Delta(\tau)|) \le -\alpha_3(\|x(\tau)\|) + \alpha_4(\|\Delta\|_{L^\infty[0,t]}) \quad \text{for a.e. } \tau \in [0,t]. \tag{5.32}$$

Clearly,

$$V(x(t)) \le \big(\alpha_2 \circ \alpha_3^{-1}\big)(2\alpha_4(\|\Delta\|_{L^\infty[0,t]})) \implies \|x(t)\| \le \beta_2(\|\Delta\|_{L^\infty[0,t]}).$$

Moreover,

$$V(x(t)) > \big(\alpha_2 \circ \alpha_3^{-1}\big)(2\alpha_4(\|\Delta\|_{L^\infty[0,t]})) \implies \alpha_3(\|x(t)\|) > 2\alpha_4(\|\Delta\|_{L^\infty[0,t]}),$$

which, together with (5.32), yields

$$\begin{aligned} &V(x(t)) > \big(\alpha_2 \circ \alpha_3^{-1}\big)(2\alpha_4(\|\Delta\|_{L^\infty[0,t]})) \\ &\quad\implies (V \circ x)'(\tau) < -\tfrac{1}{2}\alpha_3(\|x(\tau)\|) \le -\tfrac{1}{2}\alpha_5(V(x(\tau))) \le -\tfrac{1}{2}\alpha_7(V(x(\tau))) \\ &\qquad\qquad = Z(V(x(\tau))) \quad \text{for a.e. } \tau \in [0,t], \end{aligned}$$

and so

$$\begin{aligned} &V(x(t)) > \big(\alpha_2 \circ \alpha_3^{-1}\big)(2\alpha_4(\|\Delta\|_{L^\infty[0,t]}) \\ &\qquad \implies V(x(t)) \le \beta_0(t, V(x^0)) \implies \|x(t)\| \le \beta_1(t, \|x^0\|). \end{aligned}$$

Therefore, $\|x(t)\| \le \max\big\{\beta_1(t, \|x^0\|)\,,\, \beta_2(\|\Delta\|_{L^\infty[0,t]})\big\}\ \forall\ t \in \mathbb{R}_+$. □

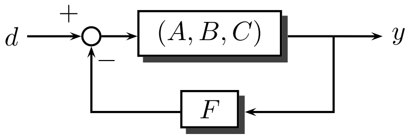


FIG. 6.1. *Interconnection of linear system (A, B, C) and hysteresis operator F.*

6. Hysteretic feedback systems. We return to the feedback interconnection of Figure 1.2, but now in a single-input ($t \mapsto d(t) \in \mathbb{R}$), single-output ($t \mapsto y(t) \in \mathbb{R}$) setting and with a hysteresis operator F in the feedback path, as shown in Figure 6.1. We deem an operator $F : C(\mathbb{R}_+) \to C(\mathbb{R}_+)$ to be a *hysteresis operator* if it is both causal and rate independent. By *rate independence*, we mean that $F(y \circ \zeta) = (Fy) \circ \zeta$ for every $y \in C(\mathbb{R}_+)$ and every time transformation $\zeta : \mathbb{R}_+ \to \mathbb{R}_+$ (that is, a continuous, nondecreasing, and surjective map). Conditions on F which ensure well-posedness of the feedback interconnection (existence and uniqueness of solutions of the associated initial-value problem) are expounded in, for example, [16] and [17]. Whilst, in principle, the ensuing analysis is applicable in the context of any causal operator F that can be embedded in a set-valued map $\Phi \in \mathcal{U}$, for clarity of presentation, we focus on the class of Preisach operators.

Preisach and Prandtl hysteresis. The Preisach operator described in this section encompasses both backlash and Prandtl operators. It can model complex hysteresis effects: For example, nested loops in input-output characteristics. A basic building block for these operators is the *backlash* operator. A discussion of the *backlash* operator (also called *play* operator) can be found in a number of references; see, for example, [5], [14], and [15]. Let $\sigma \in \mathbb{R}_+$ and introduce the function $b_\sigma : \mathbb{R}^2 \to \mathbb{R}$ given by

$$b_\sigma(v_1, v_2) := \max\{v_1 - \sigma\, ,\, \min\{v_1 + \sigma, v_2\}\} = \begin{cases} v_1 - \sigma, & \text{if } v_2 < v_1 - \sigma, \\ v_2, & \text{if } v_2 \in [v_1 - \sigma, v_1 + \sigma], \\ v_1 + \sigma, & \text{if } v_2 > v_1 + \sigma. \end{cases}$$

Let $C_{\mathrm{pm}}(\mathbb{R}_+)$ denote the space of continuous piecewise monotone functions defined on $\mathbb{R}_+$. For all $\sigma \in \mathbb{R}_+$ and $\zeta \in \mathbb{R}$, define the operator $\mathcal{B}_{\sigma,\,\zeta} : C_{\mathrm{pm}}(\mathbb{R}_+) \to C(\mathbb{R}_+)$ by

$$\mathcal{B}_{\sigma,\,\zeta}(y)(t) = \begin{cases} b_\sigma(y(0), \zeta) & \text{for } t = 0\,, \\ b_\sigma(y(t), (\mathcal{B}_{\sigma,\,\zeta}(u))(t_i)) & \text{for } t_i < t \le t_{i+1},\ i = 0, 1, 2, \dots, \end{cases}$$

where $0 = t_0 < t_1 < t_2 < \dots$, $\lim_{n\to\infty} t_n = \infty$, and u is monotone on each interval $[t_i, t_{i+1}]$. We remark that ζ plays the role of an "initial state." It is not difficult to show that the definition is independent of the choice of the partition (t_i). Figure 6.2 illustrates how $\mathcal{B}_{\sigma,\,\zeta}$ acts. It is well known that $\mathcal{B}_{\sigma,\,\zeta}$ extends to a Lipschitz continuous hysteresis operator on $C(\mathbb{R}_+)$ (with Lipschitz constant $L = 1$), the so-called backlash operator, which we shall denote by the same symbol $\mathcal{B}_{\sigma,\,\zeta}$.

Let $\xi : \mathbb{R}_+ \to \mathbb{R}$ be a compactly supported and globally Lipschitz function with Lipschitz constant 1. Let μ be a regular signed Borel measure on $\mathbb{R}_+$. Denoting Lebesgue measure on $\mathbb{R}$ by μ_L, let $w : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ be a locally $(\mu_L \otimes \mu)$-integrable function, and let $w_0 \in \mathbb{R}$. The operator $\mathcal{P}_\xi : C(\mathbb{R}_+) \to C(\mathbb{R}_+)$ defined by

$$(6.1) \quad (\mathcal{P}_\xi(y))(t) = \int_0^\infty \int_0^{(\mathcal{B}_{\sigma,\,\xi(\sigma)}(y))(t)} w(s,\sigma)\mu_L(\mathrm{d}s)\mu(\mathrm{d}\sigma) + w_0 \qquad \forall\, u \in C(\mathbb{R}_+) \ \ \forall\, t \in \mathbb{R}_+$$

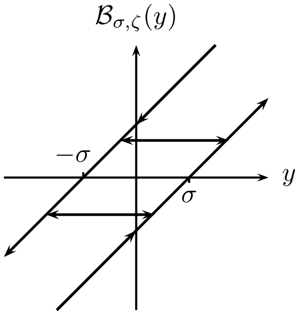


FIG. 6.2. *Backlash hysteresis.*

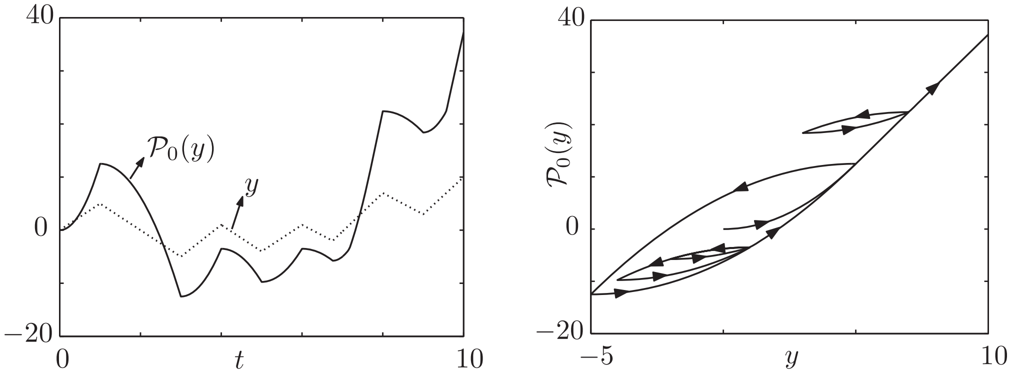


FIG. 6.3. *Example of Prandtl hysteresis.*

is called a *Preisach* operator: This definition is equivalent to that adopted in [5, section 2.4]. It is well known that $\mathcal{P}_\xi$ is a hysteresis operator (this follows from the fact that $\mathcal{B}_{\sigma,\,\xi(\sigma)}$ is a hysteresis operator for every $\sigma \geq 0$). Under the assumption that the measure μ is finite and w is essentially bounded, the operator $\mathcal{P}_\xi$ is Lipschitz continuous with Lipschitz constant $L = |\mu|(\mathbb{R}_+)\|w\|_\infty$ (see [15]) in the sense that

$$\sup_{t\in\mathbb{R}_+} |\mathcal{P}_\xi(y_1)(t) - \mathcal{P}_\xi(y_2)(t)| \leq L \sup_{t\in\mathbb{R}_+} |y_1(t) - y_2(t)| \quad \forall\, y_1, y_2 \in C(\mathbb{R}_+).$$

This property ensures the well-posedness of the feedback interconnection.

Setting $w(\cdot,\cdot) = 1$ and $w_0 = 0$ in (6.1), we obtain the *Prandtl* operator $\mathcal{P}_\xi : C(\mathbb{R}_+) \to C(\mathbb{R}_+)$ defined by

$$\mathcal{P}_\xi(y)(t) = \int_0^\infty (\mathcal{B}_{\sigma,\,\xi(\sigma)}(y))(t)\mu(\mathrm{d}\sigma) \quad \forall\, u \in C(\mathbb{R}_+) \;\; \forall\, t \in \mathbb{R}_+. \tag{6.2}$$

For $\xi \equiv 0$ and μ given by $\mu(E) = \int_E \chi_{[0,5]}(\sigma)\mathrm{d}\sigma$ (where $\chi_{[0,5]}$ denotes the indicator function of the interval $[0,5]$), the Prandtl operator is illustrated in Figure 6.3.

The next proposition identifies conditions under which Preisach operator (6.1) satisfies a generalized sector bound. For simplicity, we assume that the measure μ and the function w are nonnegative (an important case in applications), although the proposition can be extended to signed measures μ and sign-indefinite functions w.

Proposition 6.1. *Let $\mathcal{P}_\xi$ be the Preisach operator defined in* (6.1). *Assume that the measure μ is nonnegative, $a_1 := \mu(\mathbb{R}_+) < \infty$, $a_2 := \int_0^\infty \sigma\mu(\mathrm{d}\sigma) < \infty$, $b_1 := \operatorname{ess\,inf}_{(s,\sigma)\in\mathbb{R}\times\mathbb{R}_+} w(s,\sigma) \geq 0$, $b_2 := \operatorname{ess\,sup}_{(s,\sigma)\in\mathbb{R}\times\mathbb{R}_+} w(s,\sigma) < \infty$, and set*

$$a_\mathcal{P} := a_1 b_1, \quad b_\mathcal{P} := a_1 b_2\,, \quad c_\mathcal{P} := a_2 b_2 + |w_0|. \tag{6.3}$$

Then, for all $y \in C(\mathbb{R}_+)$ and all $t \in \mathbb{R}_+$,

$$y(t) \geq 0 \implies a_\mathcal{P} y(t) - c_\mathcal{P} \leq (\mathcal{P}_\xi(y))(t) \leq b_\mathcal{P} y(t) + c_\mathcal{P}, \tag{6.4}$$

$$y(t) \leq 0 \implies b_\mathcal{P} y(t) - c_\mathcal{P} \leq (\mathcal{P}_\xi(y))(t) \leq a_\mathcal{P} y(t) + c_\mathcal{P}, \tag{6.5}$$

and furthermore, for every $\eta > 0$,

$$|y(t)| \geq c_\mathcal{P}/\eta \implies (a_\mathcal{P} - \eta) y^2(t) \leq (\mathcal{P}_\xi(y))(t) y(t) \leq (b_\mathcal{P} + \eta) y^2(t). \tag{6.6}$$

Proof. Let $y \in C(\mathbb{R}_+)$ and $t \in \mathbb{R}_+$ be arbitrary. Note initially that, by the definition of the backlash operator,

$$\big(\mathcal{B}_{\sigma,\xi(\sigma)}(y)\big)(t) \in [y(t) - \sigma, y(t) + \sigma] \quad \forall\, \sigma \in \mathbb{R}_+.$$

Case 1. Assume $y(t) \geq 0$. Writing $E_1 := [0, y(t)]$ and $E_2 := (y(t), \infty)$, we have

$$\begin{aligned}
\big(\mathcal{P}_\xi y\big)(t) &\geq \left(\int_{E_1} + \int_{E_2}\right) \int_0^{y(t)-\sigma} w(s,\sigma)\mu_L(\mathrm{d}s)\mu(\mathrm{d}\sigma) - |w_0| \\
&\geq b_1 \int_{E_1} (y(t) - \sigma)\mu(\mathrm{d}\sigma) + b_2 \int_{E_2} (y(t) - \sigma)\mu(\mathrm{d}\sigma) - |w_0| \\
&= \big(b_1\mu(E_1) + b_2\mu(E_2)\big)y(t) - b_1 \int_{E_1} \sigma\,\mu(\mathrm{d}\sigma) - b_2 \int_{E_2} \sigma\,\mu(\mathrm{d}\sigma) - |w_0| \\
&\geq a_1 b_1 y(t) - a_2 b_2 - |w_0| = a_\mathcal{P} y(t) - c_\mathcal{P}.
\end{aligned}$$

Moreover,

$$\begin{aligned}
\big(\mathcal{P}_\xi y\big)(t) &\leq \int_0^\infty \int_0^{y(t)+\sigma} w(s,\sigma)\mu_L(\mathrm{d}s)\mu(\mathrm{d}\sigma) + |w_0| \\
&\leq b_2 \int_0^\infty (y(t) + \sigma)\mu(\mathrm{d}\sigma) + |w_0| \leq a_1 b_2 y(t) + a_2 b_2 + |w^0| = b_\mathcal{P} y(t) + c_\mathcal{P}.
\end{aligned}$$

This establishes (6.4).

Case 2. Now assume $y(t) \leq 0$. The argument used in Case 1 applies mutatis mutandis to conclude (6.5).

Finally, inequality (6.6) is a straightforward consequence of (6.4) and(6.5). □

For example, the Prandtl operator in Figure 6.3 satisfies the hypotheses of Proposition 6.1.

Let $\mathcal{P}_\xi$ be a Preisach operator satisfying the hypotheses of Proposition 6.1. Let $a_\mathcal{P}$, $b_\mathcal{P}$, and $c_\mathcal{P}$ be given by (6.3) and define Φ, $\tilde{\Phi} \in \mathcal{U}$ by

$$\Phi(y) := \begin{cases} \{v \in \mathbb{R} \mid a_\mathcal{P} y - c_\mathcal{P} \leq v \leq b_\mathcal{P} y + c_\mathcal{P}\}, & y \geq 0, \\ \{v \in \mathbb{R} \mid b_\mathcal{P} y - c_\mathcal{P} \leq v \leq a_\mathcal{P} y + c_\mathcal{P}\}, & y < 0. \end{cases}$$

$$\tilde{\Phi}(y) := \{v \in \mathbb{R} \mid (a_\mathcal{P} - \eta) y^2 \leq vy \leq (b_\mathcal{P} + \eta) y^2\},$$

where $\eta > 0$. In view of (6.4) and (6.5),

$$y \in C(\mathbb{R}_+) \quad \Longrightarrow \quad (\mathcal{P}_\xi(y))(t) \in \Phi(y(t)) \ \ \forall\, t \in \mathbb{R}_+.$$

Moreover, writing $K := [-c_{\mathcal{P}}/\eta\,,\, c_{\mathcal{P}}/\eta]$, we have

$$\Phi(y) \subset \tilde{\Phi}(y) \quad \forall\, y \in \mathbb{R}\backslash K \quad \text{and} \quad E := \sup_{y\in K} \sup_{v\in\Phi(y)} \inf_{\tilde{v}\in\tilde{\Phi}(y)} |v - \tilde{v}| = c_{\mathcal{P}}.$$

Let linear system (A, B, C) (with transfer function G) be stabilizable and detectable. Write $a := a_{\mathcal{P}} - \eta$, $b := b_{\mathcal{P}} + \eta$, and assume that $G/(1 + aG) \in H^\infty$, and for some $\delta \in (0, 1)$, $(1 + bG)/(1 + aG) - \delta$ is positive real. Then hypothesis (H1) holds with $m = 1$ and $\tilde{\Phi}$ replacing Φ.

Example. As a concrete example, consider a mechanical system with damping coefficient $\gamma > 0$ and a hysteretic restoring force in the form of backlash, with real parameters $\sigma > 0$ and ζ:

$$\ddot{y}(t) + \gamma\dot{y}(t) + \mathcal{B}_{\sigma,\zeta}(y)(t) = d(t). \tag{6.7}$$

Setting $w(\cdot,\cdot) := 1$, $w^0 = 0$, $\mu := \delta_\sigma$ (the Dirac measure with support $\{\sigma\}$), and $\xi(\cdot) := \zeta$ in (6.1), we see that $\mathcal{B}_{\sigma,\zeta} = \mathcal{P}_\xi$. In this case and in the notation of Proposition 6.1, we have $a_1 = b_1 = b_2 = a_{\mathcal{P}} = b_{\mathcal{P}} = 1$ and $a_2 = c_{\mathcal{P}} = \sigma$. Choosing $\eta \in (0, 1)$, we have $0 < a < b$, where, as before, $a = a_{\mathcal{P}} - \eta$ and $b = b_{\mathcal{P}} + \eta$ and by Proposition 6.1,

$$|y(t)| \geq \sigma/\eta \implies ay^2(t) \leq \big(\mathcal{B}_{\sigma,\xi}(y)\big)(t)y(t) \leq by^2(t).$$

The transfer function G is given by $G(s) = 1/(s^2 + \gamma s)$, $G/(1 + aG)$ is given by $1/(s^2 + \gamma s + a)$, and $(1 + bG)/(1 + aG) - \delta$ is given by $(1 - \delta) + 2\eta/(s^2 + \gamma s + a)$. Clearly, $G/(1+aG) \in H^\infty$ and a straightforward calculation reveals that, for all $\eta > 0$ sufficiently small, $(1 + bG)/(1 + aG) - \delta$ is positive real.

Returning to the general setting, we are now in a position to invoke Corollary 3.6 to conclude properties of solutions of the single-input, single-output functional differential equation

$$\dot{x}(t) = Ax(t) + B\big[d(t) - (\mathcal{P}_\xi(Cx))(t)\big], \quad x(0) = x^0. \tag{6.8}$$

We reiterate that, for each $x^0 \in \mathbb{R}^n$ and $d \in L^\infty_{\text{loc}}(\mathbb{R}_+)$, (6.8) has a unique global solution. An application of Corollary 3.6 (with $\Delta(t) = \{d(t)\}$ for all $t \in \mathbb{R}_+$) yields the existence of constants $\varepsilon, c_1, c_2 > 0$ such that, for every global solution x,

$$\|x(t)\| \leq c_1 e^{-\varepsilon t}\|x^0\| + c_2\left(\|d\|_{L^\infty[0,t]} + c_{\mathcal{P}}\right) \quad \forall\, t \in \mathbb{R}_+, \tag{6.9}$$

showing, in particular, that (6.8) is input-to-state stable with bias $c_{\mathcal{P}}$. Furthermore, by Lemma 3.3,

$$\lim_{t\to\infty} d(t) = 0 \quad \Longrightarrow \quad \limsup_{t\to\infty} \|x(t)\| \leq c_2 c_{\mathcal{P}}. \tag{6.10}$$

We emphasize that the convergence $d(t) \to 0$ as $t \to \infty$ does, in general, not imply convergence of $x(t)$ as $t \to \infty$. To see this, consider again mechanical example (6.7). Then, for every $\gamma > 0$, there exist constants $\varepsilon, c_1, c_2 > 0$ such that (6.9) and (6.10) hold (with $x(t) = (y(t), \dot{y}(t))$ and $c_{\mathcal{P}} = \sigma$). However, we know from [17, Example 4.8] that if $d = 0$ and $\gamma \in (1, 2)$, then for all initial conditions, $\limsup_{t\to\infty} y(t) = \sigma$ and $\liminf_{t\to\infty} y(t) = -\sigma$ (equivalently, y has ω-limit set $[-\sigma, \sigma]$), showing, in particular, that $x(t) = (y(t), \dot{y}(t))$ does not converge as $t \to \infty$.

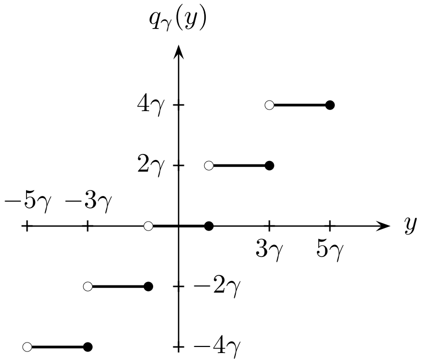


FIG. 7.1. *Uniform quantizer.*

7. Quantized feedback systems. Let (A, B, C) be a minimal realization of a linear, single-input, single-output system with transfer function G. Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous static nonlinearity with the following property:

(Q1) There exist $\varphi \in \mathcal{K}_\infty$ and a number $b > 0$ such that

$$\varphi(|y|)|y| \leq f(y)y \leq by^2 \quad \forall\, y \in \mathbb{R}.$$

Furthermore, we impose the following assumption:

(Q2) There exists $\kappa \in [0\,,\, 1/b)$ such that $\kappa + G$ is positive real.

From (Q1) and (Q2), it follows that (H3) holds with $\Phi(y) = \{f(y)\}$ and $\delta = \kappa b \in [0, 1)$. Consequently, by Theorem 3.5, the system

$$\dot{x}(t) = Ax(t) + B(d(t) - f(Cx(t))), \quad x(0) = x^0 \in \mathbb{R}^n, \quad d \in L^\infty_{\mathrm{loc}}(\mathbb{R}_+) \tag{7.1}$$

is input-to-state stable. Now consider (7.1) subject to quantization of the output $y = Cx$, that is, the system

$$\dot{x}(t) = Ax(t) + B\big(d(t) - (f \circ q_\gamma)(Cx(t))\big), \quad x(0) = x^0 \in \mathbb{R}^n, \quad d \in L^\infty_{\mathrm{loc}}(\mathbb{R}_+), \tag{7.2}$$

where $q_\gamma : \mathbb{R} \to \mathbb{R}$, parameterized by $\gamma > 0$, is a uniform quantizer (see Figure 7.1) given by

$$q_\gamma(y) = 2(m+1)\gamma \quad \forall\, y \in \big((2m+1)\gamma,\, (2m+3)\gamma\big] \quad \forall\, m \in \mathbb{Z}.$$

We interpret the differential equation (with discontinuous right-hand side) in (7.1) in a set-valued sense by embedding the quantizer q_γ in the set-valued map $Q_\gamma \in \mathcal{U}$ defined by

$$Q_\gamma(y) := \begin{cases} \{q_\gamma(y)\}, & y \in \big((2m+1)\gamma\,,\, (2m+3)\gamma\big), \quad m \in \mathbb{Z}, \\ [2m\gamma\,,\, 2(m+1)\gamma], & y = (2m+1)\gamma, \quad m \in \mathbb{Z}, \end{cases}$$

and subsuming (7.2) in the differential inclusion

$$\dot{x}(t) - Ax(t) \in B\big(\Delta(t) - \Phi_\gamma(Cx(t))\big), \quad x(0) = x^0 \in \mathbb{R}^n, \quad \Delta \in \mathcal{B}, \tag{7.3}$$

where $\Delta : t \mapsto \{d(t)\}$ and $\Phi_\gamma \in \mathcal{U}$ is given by

$$\Phi_\gamma(y) := f(Q_\gamma(y)) = \{f(\xi) \mid \ \xi \in Q_\gamma(y)\}.$$

Choose $\varepsilon \in (0,1)$ sufficiently small so that $(1+\varepsilon)\kappa < 1/b$. Write $\tilde{b} := (1+\varepsilon)b$ and define $\tilde{\varphi} \in \mathcal{K}_\infty$ by $\tilde{\varphi}(s) := \varphi((1-\varepsilon)s) \ \forall\, s \in \mathbb{R}_+$.

Lemma 7.1. *There exists $M \in \mathbb{N}$ such that, for every $\gamma > 0$,*

$$y \in \mathbb{R},\ |y| \geq \gamma M,\ v \in \Phi_\gamma(y) \quad \Longrightarrow \quad \tilde{\varphi}(|y|)|y| \leq yv \leq \tilde{b}y^2.$$

Proof. Observe that, for all $m \in \mathbb{N}$,

$$\frac{2m+2}{2m+3} \leq \frac{w}{y} \leq \frac{2m+4}{2m+1} \quad \forall\, w \in Q_\gamma(y) \ \ \forall\, y \in \big((2m+1)\gamma,\, (2m+3)\gamma\big].$$

Therefore, there exists $M \in \mathbb{N}$ such that

$$(1-\varepsilon)y^2 \leq wy \leq (1+\varepsilon)y^2 \quad \forall\, w \ \in Q_\gamma(y) \ \ \forall\, y \geq \gamma M.$$

Since Q_γ has odd symmetry ($Q_\gamma(y) = -Q_\gamma(-y)$), it immediately follows that

$$(1-\varepsilon)y^2 \leq wy \leq (1+\varepsilon)y^2 \quad \forall\, w \ \in Q_\gamma(y) \ \ \forall\, |y| \geq \gamma M. \tag{7.4}$$

Let y be such that $|y| \geq \gamma M$, and let $v \in \Phi_\gamma(y)$. Then $v = f(w)$ for some $w \in Q_\gamma(y)$. Invoking (Q1) and (7.4), it follows that

$$\varphi(|w|)|y| = \varphi(|w|)|w|\frac{y}{w} \leq f(w)w\frac{y}{w} = f(w)y = vy \leq bwy \leq (1+\varepsilon)by^2 = \tilde{b}y^2. \tag{7.5}$$

Since $\varphi(|w|) = \varphi(|w||y|/|y|) = \varphi(wy/|y|)$ and invoking (7.4) and (7.5), we have

$$\tilde{\varphi}(|y|)|y| = \varphi((1-\varepsilon)|y|)|y| \leq \varphi(|w|)|y| \leq vy \leq \tilde{b}y^2.$$

This completes the proof. □

Let $M \in \mathbb{N}$ be as in Lemma 7.1 and define $\tilde{\Phi} \in \mathcal{U}$ by

$$\tilde{\Phi}(y) := \begin{cases} [\tilde{\varphi}(|y|)\,,\, \tilde{b}|y|], & y \geq 0, \\ [-\tilde{b}|y|\,,\, -\tilde{\varphi}(|y|)], & y < 0. \end{cases} \tag{7.6}$$

Clearly,

$$y \in \mathbb{R},\ v \in \tilde{\Phi}(y) \quad \Longrightarrow \quad \max\big\{\tilde{\varphi}(|y|)|y|,\, v^2/\tilde{b}\big\} \leq yv,$$

and, by Lemma 7.1, we also have $\Phi_\gamma(y) \subset \tilde{\Phi}(y) \ \forall\, y \in \mathbb{R} \setminus [-\gamma M, \gamma M]$. Moreover, by (Q2) (and recalling that $\kappa\tilde{b} < 1$), $(\delta/\tilde{b}) + G$ is positive real for every $\delta \in [\kappa\tilde{b}\,, 1)$. We are now in a position to invoke Corollary 3.7 (with $K = [-\gamma M, \gamma M]$) to conclude the existence of $\beta_1 \in \mathcal{KL}$ and $\beta_2 \in \mathcal{K}_\infty$, which do not depend on $\gamma > 0$ (recall Remark 3.8) such that, for all $\gamma > 0$, all $x^0 \in \mathbb{R}^n$, and all $d \in L^\infty_{\mathrm{loc}}(\mathbb{R}_+)$, every global solution of (7.3), with $\Delta : t \mapsto \{d(t)\}$ satisfies

$$\|x(t)\| \leq \max\big\{\beta_1(t, \|x^0\|)\,,\, \beta_2(\|d\|_{L^\infty[0,t]} + E_\gamma)\big\} \quad \forall\, t \in \mathbb{R}_+,$$

where $E_\gamma := \sup_{|y|\le\gamma M}\sup_{v\in\Phi_\gamma(y)}\inf_{\tilde v\in\tilde\Phi_\gamma(y)}|v-\tilde v|$. Noting that $E_\gamma \to 0$ as $\gamma \downarrow 0$ (if f is locally Lipschitz, then $E_\gamma = O(\gamma)$ as $\gamma \downarrow 0$), we may conclude robustness with respect to quantization in the sense that the quantized feedback system is such that the unbiased ISS property of unquantized system (7.1) is approached as $\gamma \downarrow 0$.

8. Conclusion. Feedback interconnections consisting of a linear system in the forward path and a nonlinearity in the feedback path have been considered. Adopting a differential inclusions framework, nonlinearities of considerable generality are encompassed, including inter alia both hysteresis operators and quantization operators. Conditions on the linear and nonlinear components have been identified (in Theorems 3.4 and 3.5) under which ISS (and a fortiori global asymptotic stability of the zero state) of the feedback interconnection is assured. The results of this paper are in the spirit of absolute stability theory: in particular, when specialized appropriately, classical absolute stability results pertaining to the circle criterion are recovered. In Corollaries 3.6 and 3.7, hypotheses are imposed on the nonlinearities (namely, generalized sector conditions) considerably weaker than those posited in Theorems 3.4 and 3.5, under which ISS with bias (and a fortiori asymptotic stability of a compact neighborhood of the zero state) may be concluded. Applications of the results to systems with hysteresis and to systems with output quantization have been detailed.

REFERENCES

[1] M. Arcak and A. Teel, *Input-to-state stability for a class of Lurie systems*, Automatica, 38 (2002), pp. 1945–1949.

[2] J.P. Aubin and A. Cellina, *Differential Inclusions*, Springer-Verlag, Berlin, 1984.

[3] N.A. Barabanov and V.A.Yakubovich, *Absolute stability of control systems with one hysteresis nonlinearity*, Autom. Remote Control, 12 (1979), pp. 5–12.

[4] R.W. Brockett and D. Liberzon, *Quantized feedback stabilization of linear systems*, IEEE Trans. Automat. Control, 45 (2000), pp. 1279–1289.

[5] M. Brokate and J. Sprekels, *Hysteresis and Phase Transitions*, Springer-Verlag, New York, 1996.

[6] K. Deimling, *Multivalued Differential Equations*, de Gruyter, Berlin, 1992.

[7] C.A. Desoer and M. Vidyasagar, *Feedback Systems: Input-Output Properties*, Academic Press, New York, 1975.

[8] M. Fu and L. Xie, *The sector bound approach to quantized feedback control*, IEEE Trans. Automat. Control, 50 (2005), pp. 1698–1711.

[9] R.B. Gorbet, K.A. Morris, and D.W.L. Wang, *Passivity-based stability and control of hysteresis in smart actuators*, IEEE Trans. Control Syst. Technol., 9 (2001), pp. 5–16.

[10] W. Hahn, *Stability of Motion*, Springer-Verlag, Berlin, 1967.

[11] B. Jayawardhana, H. Logemann, and E.P. Ryan, *Infinite-dimensional feedback systems: The circle criterion and input-to-state stability*, Commun. Inf. Syst., to appear.

[12] U. Jönson, *Stability of uncertain systems with hysteresis nonlinearities.* Internat. J. Robust Nonlinear Control, 8 (1998), pp. 279–293.

[13] H.K. Khalil, *Nonlinear Systems*, 3rd ed., Prentice-Hall, Upper Saddle River, NJ, 2002.

[14] M.A. Krasnosel'skii and A.V. Pokrovskii, *Systems with Hysteresis*, Springer-Verlag, Berlin, 1989.

[15] H. Logemann and A.D. Mawby, *Low-gain integral control of infinite-dimensional regular linear systems subject to input hysteresis*, in Advances in Mathematical Systems Theory, F. Colonius et al., eds., Birkhäuser, Boston, 2001, pp. 255–293.

[16] H. Logemann and E.P. Ryan, *Systems with hysteresis in the feedback loop: Existence, regularity and asymptotic behavior of solutions*, ESAIM Control, Optim. Calc. Var., 9 (2003), pp. 169–196.

[17] H. Logemann, E.P. Ryan, and I. Shvartsman, *A class of differential-delay systems with hysteresis: Asymptotic behavior of solutions*, Nonlinear Anal., 69 (2008), pp. 363–391.

[18] A.I. Lur'e and V.N. Postnikov, *On the theory of stability of control systems*, Appl. Math. Mech., 8 (1944), pp. 246–248 (in Russian).

[19] K.S. Narendra and J.H. Taylor, *Frequency Criteria for Absolute Stability*, Academic Press, New York, 1973.

[20] E.D. Sontag, *Input to state stability: Basic concepts and results*, in Nonlinear and Optimal Control Theory, P. Nistri and G. Stefani, eds., Springer-Verlag, Berlin, 2006, pp. 163–220.

[21] M. Vidyasagar, *Nonlinear Systems Analysis*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1993.

[22] R. Vinter, *Optimal Control*, Birkhäuser, Boston, 2000.

[23] V.A. Yakubovich, G.A. Leonov, and A.Kh. Gelig, *Stability of Stationary Sets in Control Systems with Discontinuous Nonlinearities*, World Scientific, Singapore, 2004.

SIAM J. Control Optim.
Vol. 48, No. 2, pp. 1055–1079

© 2009 Society for Industrial and Applied Mathematics

CONTROL SYSTEMS WITH ALMOST PERIODIC EXCITATIONS[*]

FRITZ COLONIUS[†] AND TOBIAS WICHTREY[†]

Abstract. For control systems described by ordinary differential equations subject to almost periodic excitations the controllability properties depend on the specific excitation. Here these properties and, in particular, control sets and chain control sets are discussed for all excitations in the closure of all time shifts of a given almost periodic function. Then relations between heteroclinic orbits of an uncontrolled and unperturbed system and controllability for small control ranges and small perturbations are studied using Melnikov's method. Finally, a system with two-well potential is studied in detail.

Key words. nonautonomous control systems, almost periodicity, control sets, Melnikov method

AMS subject classifications. 93B05, 37N35, 34C37

DOI. 10.1137/070704733

1. Introduction.

This paper analyzes controllability properties of control systems which are subject to almost periodic excitations. More precisely, we consider

$$\dot{x}(t) = f\left(x(t), z(t), u(t)\right), \; u \in \mathcal{U}, \tag{1.1}$$

in an open set $M \subset \mathbb{R}^d$ with admissible controls in $\mathcal{U} = \{u \in L_\infty(\mathbb{R}, \mathbb{R}^m),\; u(t) \in U$ for all $t \in \mathbb{R}\}$ and control range $U \subset \mathbb{R}^m$. We assume that z is an almost periodic function with values in a compact subset $Z \subset \mathbb{R}^k$. In particular, this includes periodic excitations and excitations with several incommensurable periods.

Instead of analyzing the behavior of system (1.1) for a single almost periodic excitation, we allow time shifts of z and, more generally, all excitations in the set $\mathcal{Z}$ of continuous functions which can uniformly be approximated by shifts of z (again, all elements of $\mathcal{Z}$ are almost periodic). Observe that the trajectories of (1.1) are determined by the initial states $x = x(0) \in M$, the excitation $z \in \mathcal{Z}$, and the control function $u : \mathbb{R} \to \mathbb{R}^m$.

There are various ways to look at this system:

(i) as a control system in M with states $x \in M$;

(ii) as a control system in $M \times \mathcal{Z}$ with extended states $(x, z) \in M \times \mathcal{Z}$;

(iii) as a dynamical system in $M \times \mathcal{Z} \times \mathcal{U}$ with states $(x, z, u) \in M \times \mathcal{Z} \times \mathcal{U}$.

Observe that the control system in (i) is nonautonomous; the evolution of the states x is determined only if, in addition to the control function $u \in \mathcal{U}$, also the phase of the almost periodic function z is known. Hence, here we have to distinguish between an analysis for fixed excitation $z \in \mathcal{Z}$ and the projections to M. In (ii), we can sometimes, if the almost periodic function is a solution of a differential equation on a compact manifold Z (e.g., if Z is a k-torus), replace $\mathcal{Z}$ by Z. Here, however, exact controllability properties in the extended state space $M \times Z$ can only hold in the very special case of a periodic function z. Furthermore, the dimension of the state space of the control system is increased by k, which makes a global numerical analysis much more difficult. The formulation (iii) results in a continuous dynamical system

[*] Received by the editors October 8, 2007; accepted for publication (in revised form) November 7, 2008; published electronically March 6, 2009.

http://www.siam.org/journals/sicon/48-2/70473.html

[†] Institut für Mathematik, Universität Augsburg, 86135 Augsburg, Germany (fritz.colonius@math.uni-augsburg.de, tobias.wichtrey@math.uni-augsburg.de).

(a control flow) provided that the system is control affine and the control range U is compact and convex. The analysis of this dynamical system (including time shifts on $\mathcal{Z}$ and on $\mathcal{U}$) may yield structural insights and, in particular, sheds light on subsets of complete controllability, i.e., control sets. In the present paper, we will analyze system (1.1) employing all three points of view above.

Note that, for T-periodically excited control systems, controllability properties in the extended state space (where also the phase in $\mathbb{R}/T\mathbb{Z}$ is part of the state) can essentially be characterized by a Poincaré section, i.e., the intersection with a fiber over a fixed phase (compare Gayer [10]). The almost periodic case considered here is more complicated and makes it necessary to use the fact that the time shift on $\mathcal{Z}$ generates a minimal flow. This will allow us to show that, in an appropriately generalized sense, control sets and chain control sets are to a large extent determined by their intersection with a single fiber over a phase $z \in \mathcal{Z}$.

Using methods from ergodic theory, controllability properties of nonautonomous linear control systems have also been discussed by Johnson and Nerurkar [13]. Many further results in this direction have been obtained, in particular, in connection with associated Riccati equations. For a different line of research, see San Martin and Patrao [19], who study control sets and chain control sets for semidynamical systems on fiber bundles (related to the third interpretation above of system (1.1)).

The main topic of this paper are the relations between hetero- or homoclinic orbits of an uncontrolled and unperturbed system and controllability for small control ranges. Here Melnikov's method plays an important role, as observed, from a numerical point of view, in Colonius et al. [5] where this method is used to analyze two models for ship dynamics (without periodic excitation). In the present paper, a characterization for systems with general almost periodic excitations will be given. Apart from [5], to the best of our knowledge, Melnikov's method has not been used in the literature to prove controllability results. Hence, our results are also new in the periodic case. Melnikov's method for differential equations with almost periodic excitations was, in particular, developed by Palmer [18], Scheurle [21], and Meyer and Sell [17]. Our paper is closer to the spirit of the latter reference, since we consider the hull of an almost periodic excitation. We would like to point out that we do not really need the strength of Melnikov's result here; existence of a chaotic set is not in our center of interest. Instead, intersections of stable and unstable manifolds are relevant here. Note that basic references for almost periodic differential equations include Fink [9] and Levitan and Zhikov [15]; a nice discussion of almost periodic and quasi-periodic functions can also be found in section II.1 of [17], together with further references.

Furthermore, we apply our results to a second order system modeling ship dynamics and capsizing under wave excitations. This system involving an M-potential has been proposed by Kreuzer and Sichermann [14] for roll motion in following seas. These models have a rich history; see, among others, Falzarano, Shaw, and Troesch [8]; Thompson [26]; Hsieh, Troesch, and Shaw [12]; Soliman and Thompson [23]; and Colonius, Gayer, and Kliemann [3]. In particular, Gayer [10] analyzed the so-called escape equation with periodic excitation interpreting the time-dependent perturbations as controls and gave a detailed analysis of the behavior under increasing perturbation ranges. His analysis has to be complemented by the global controllability behavior due to motions close to homoclinic orbits (compare also [5]).

The paper is organized as follows: After preliminaries in section 2, we analyze chain control sets in section 3. Section 4 introduces control sets and presents relations to chain control sets and to almost periodic solutions of the uncontrolled system. Section 5 presents relevant results on almost periodic perturbations of hyperbolic equilibria and Melnikov's method. These results are essentially known in the literature (see Palmer [18], Scheurle [21], and also Meyer and Sell [17]). However, for the reader's convenience, we have included some arguments from the proofs. This is used in section 6 to study the relation between heteroclinic orbits of an unperturbed system and controllability for small control ranges. Finally, in section 7 we discuss a second order system with M-potential modeling ship roll motion.

2. Preliminaries.

Consider the control system (1.1)

$$\dot{x}(t) = f\left(x(t), z(t), u(t)\right),\ u \in \mathcal{U},$$

in an open set $M \subset \mathbb{R}^d$ with admissible controls in $\mathcal{U}$, and assume that z is an almost periodic function. That is, we assume (compare, e.g., Scheurle [21, Definition 2.6]) that $z : \mathbb{R} \to \mathbb{R}^k$ is continuous and that for every $\varepsilon > 0$ there exists an $l = l(\varepsilon) > 0$ such that in any interval of length l there is a so-called *translation number* τ such that

$$\|z(t+\tau) - z(t)\| < \varepsilon \text{ for all } t \in \mathbb{R}.$$

Define θ as the time shift $(\theta_t z)(s) := z(t+s),\ s, t \in \mathbb{R}$. Let $\mathcal{Z}$ be the closure in the space $C_b(\mathbb{R}, \mathbb{R}^k)$ of bounded continuous functions of the shifts of an almost periodic function. Then $\mathcal{Z}$ is a minimal set; i.e., every trajectory is dense in $\mathcal{Z}$. Observe that for $z \in \mathcal{Z}$ it holds that $z(t) = (\theta_t z)(0)$. Assuming global existence and uniqueness, we denote by $\varphi(t, t_0, x, z, u)$ the solution of the initial value problem

$$\dot{x}(t) = f\left(x(t), z(t), u(t)\right),\ x(t_0) = x; \tag{2.1}$$

if $t_0 = 0$, we often omit this argument. The solution map of the coupled system is denoted by

$$\psi(t, x, z, u) = \left(\varphi(t, x, z, u), \theta_t z\right).$$

We assume that the set of admissible controls is given by

$$\mathcal{U} = \{u \in L_\infty(\mathbb{R}, \mathbb{R}^m),\ u(t) \in U \text{ for almost all } t\},$$

where $U \subset \mathbb{R}^m$. If we denote also the time shift on $\mathcal{U}$ by θ_t, we obtain the cocycle property

$$\varphi(t+s, x, z, u) = \varphi\left(s, \varphi(t, x, z, u), \theta_t z, \theta_t u\right), t, s \in \mathbb{R}.$$

Finally, the maps

$$\Phi_t : M \times \mathcal{Z} \times \mathcal{U} \to M \times \mathcal{Z} \times \mathcal{U},\ \Phi_t(x, z, u) = \left(\psi(t, x, z, u), \theta_t u\right), t \in \mathbb{R},$$

define a continuous flow, the *control flow*, provided that $U \subset \mathbb{R}^m$ is convex and compact and

$$f(x, z, u) = f_0(x, z) + \sum_{i=1}^{m} u_i f_i(x, z)$$

with C^1-functions $f_i : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}^d$; here $\mathcal{U} \subset L_\infty(\mathbb{R}, \mathbb{R}^m)$ is endowed with the weak star (weak*) topology. This follows by a minor extension of Proposition 4.1.1 in [4].

The weak* topology on $\mathcal{U}$ is compact and metrizable. Throughout this paper, we endow $\mathcal{U}$ with a corresponding metric and assume that the conditions above guaranteeing continuity of the control flow are satisfied. Note that the space $\mathcal{Z}$ of almost periodic excitations is considered in the norm topology of $C_b(\mathbb{R},\mathbb{R}^k)$. The shifts on each of these spaces are continuous.

For convenience, we also assume that $0 \in U$, and we call the corresponding differential equation with $u \equiv 0$ the uncontrolled system. For periodic and for quasi-periodic excitations we may be able to replace $\mathcal{Z}$ by a finite dimensional state space Z.

Example 2.1. For a smooth periodic excitation let $\zeta : \mathbb{S}^1 \to \mathbb{S}^1 =: Z$ be the solution map $\zeta_t z_0 = \omega(t+z_0), t \in \mathbb{R}$, of $\dot z = \omega,\ z(0) = z_0$; here $\omega > 0$ is the frequency and (2.1) may be written as

$$\dot x(t) = f\left(x(t), \zeta_t(z_0), u(t)\right),\ x(0) = x_0.$$

For a quasi-periodic excitation, let $\zeta : \mathbb{S}^k \to \mathbb{S}^k =: Z$ be the solution map $\zeta_t z_0 = (\omega_1(t+z_{0,1}), \dots, \omega_k(t+z_{0,k})), t \in \mathbb{R}$, of

$$\dot z_1 = \omega_1, \dot z_2 = \omega_2, \dots, \dot z_k = \omega_k,$$

with initial condition $z(0) = (z_{0,1}, \dots, z_{0,k})$. Here $\omega_1, \dots, \omega_k > 0$ are the frequencies and we assume that they are rationally independent; i.e., if $q_i \in \mathbb{Q}$ with $q_1\omega_1 + \cdots + q_k\omega_k = 0$, then $q_i = 0$ for all i. Again, (2.1) may be written as above.

3. Chain control sets.

In this section, we define and characterize chain control sets relative to a subset of the state space working in the general almost periodic case.

It will be convenient to write for a subset $A \subset M \times \mathcal{Z}$ the section with a fiber over $z \in \mathcal{Z}$ as

$$A_z := A \cap (M \times \{z\}).$$

Hence, $A = \bigcup_{z\in\mathcal{Z}} A_z$. Where convenient, we identify A_z and $\{x \in M,\ (x,z) \in A_z\}$.

A controlled (ε, T)-chain along $z \in \mathcal{Z}$ is given by $T_0, \dots, T_{n-1} \geq T$, controls $u_0, \dots, u_{n-1} \in \mathcal{U}$, and points $x_0, \dots, x_n \in M$ with

$$\mathrm{d}\left(\varphi(T_j, x_j, \theta_{T_0+\cdots+T_{j-1}}z, u_j), x_{j+1}\right) < \varepsilon \text{ for all } j = 0, \dots, n-1.$$

Definition 3.1. *A chain control set relative to a closed set $Q \subset M \times \mathcal{Z}$ is a nonvoid maximal set $E \subset M \times \mathcal{Z}$ such that*

(i) *for all $(x,z), (y,w) \in E$ and all $\varepsilon, T > 0$ there exists a controlled (ε, T)-chain in Q along z from x to (y,w), i.e., $x_0 = x$, $x_n = y$, and $\mathrm{d}(\theta_{T_0+\cdots+T_{n-1}}z, w) < \varepsilon$, and*

$$\psi(t, x_j, \theta_{T_0+\cdots+T_{j-1}}z, u_j) \in Q \text{ for all } t \in [0, T_j] \text{ and for all } j; \tag{3.1}$$

(ii) *for all $(x,z) \in E$ there is $u \in \mathcal{U}$ with $\psi(t,x,z,u) \in E$ for all $t \in \mathbb{R}$.*

The condition in (3.1) can be written as

$$\varphi(t, x_j, \theta_{T_0+\cdots+T_{j-1}}z, u_j) \in Q_{\theta_t z_j}.$$

Note that the three components x, z, and u are treated in different ways: Jumps are allowed in x, approximate reachability is required for z, and no condition on the controls is imposed. Observe that also Meyer and Sell [17] do not allow jumps in the almost periodic base flow. It is easy to show that chain control sets are closed.

Next, we discuss the behavior for fixed "phases" $z \in \mathcal{Z}$ by looking at the fibers of a chain control set.

LEMMA 3.2. *Suppose that E is a chain control set relative to Q. Then the fibers $E_z := E \cap Q_z, z \in \mathcal{Z}$, satisfy the following properties:*

(i) *For every $z \in \mathcal{Z}$, all $x, y \in E_z$, and all $\varepsilon, T > 0$ there exists a controlled (ε, T)-chain in Q from x along z to (y, z).*

(ii) *For every $z \in \mathcal{Z}$ and every $x \in E_z$ there exists a control $u \in \mathcal{U}$ such that*

$$\varphi(t, x, z, u) \in E_{\theta_t z} \textit{ for all } t \in \mathbb{R}.$$

(iii) *If $x_n \in E_{z_n}$ with $(x_n, z_n) \to (x, z) \in M \times \mathcal{Z}$, then $x \in E_z$.*

Proof. Condition (iii) follows from closedness of E; (i) and (ii) are obvious. □

Remark 3.3. In condition (ii) of Lemma 3.2 one does not have that a trajectory exists which after an appropriate time comes back to E_z (as for periodic excitations, where one comes back into the same fiber after the period). In the general, almost periodic case, the trajectory will never come back to the same fiber. Instead, the weaker property formulated in (ii) holds together with condition (iii), which locally connects different fibers and is an upper semicontinuity property of $z \mapsto E_z$.

Next, we discuss if the properties formulated in Lemma 3.2 characterize chain control sets.

LEMMA 3.4. *Suppose Q is compact and that $E^z \subset Q_z, z \in \mathcal{Z}$, is a family of sets satisfying conditions* (i), (ii), *and* (iii) *in Lemma* 3.2. *Assume that*

$$E := \bigcup_{z \in \mathcal{Z}} E^z \subset \operatorname{int} Q.$$

Then E satisfies properties (i) *and* (ii) *of chain control sets in Definition* 3.1.

Proof. Let $(x, z), (y, w) \in E$ and $\varepsilon, T > 0$. Then $\mathcal{Z} = \omega(z) := \{z' \in \mathcal{Z},\ \theta_{t_k} z \to z'$ for a sequence $t_k \to \infty\}$ and there is a control $u \in \mathcal{U}$ such that $\psi(t, x, z, u) \in E$ for all $t \in \mathbb{R}$. In particular, this proves property (ii) of chain control sets. Furthermore, there are $S_k > T$ such that for $z_k := \theta_{S_k} z$ one has $\mathrm{d}(z_k, w) < 1/k$ and clearly $y_k := \varphi(S_k, x, z, u) \in E_{z_k}$. By compactness of Q, we may assume that (y_k, z_k) converges to some $(y_0, w) \in Q$. By property (iii), it follows that $y_0 \in E_w$. By property (i), there is a controlled $(\varepsilon/2, T)$-chain in Q from y_0 along w to (y, w) satisfying $x_0 = y_0$, $x_n = y$, and $\mathrm{d}(\theta_{T_0+\cdots+T_{n-1}} w, w) < \varepsilon/2$, and

$$\psi(t, x_j, \theta_{T_0+\cdots+T_{j-1}} w, u_j) \in Q \text{ for all } t \in [0, T_j] \text{ and for all } j.$$

Introducing, if necessary, trivial jumps, we may assume that $T_j \in [T, 2T]$ for all j. By uniform continuity, there is $\delta > 0$ such that for all $x \in Q$ and all $u \in \mathcal{U}$

$$\mathrm{d}(z, z') < \delta \text{ implies } \mathrm{d}\left(\varphi(t, x, z, u), \varphi(t, x, z', u)\right) < \varepsilon/2,\ t \in [0, 2T]. \tag{3.2}$$

Choose k large enough such that

$$\mathrm{d}(z_k, w) = \mathrm{d}(\theta_{S_k} z, w) := \sup_{t \in \mathbb{R}} \|z(S_k + t) - w(t)\| < \delta \text{ and } \mathrm{d}(\varphi(S_k, x, z, u), y_0) < \varepsilon.$$

Hence, for all j

$$\begin{aligned} &\mathrm{d}\left(\varphi(T_j, x_j, \theta\hat{}_{S_k+T_0+\cdots+T_{j-1}} z, u_j), x_{j+1}\right) \\ &\le \mathrm{d}\left(\varphi(T_j, x_j, \theta_{S_k+T_0+\cdots+T_{j-1}} z, u_j), \varphi(T_j, x_j, \theta_{T_0+\cdots+T_{j-1}} w, u_j)\right) \\ &\quad + \mathrm{d}\left(\varphi(T_j, x_j, \theta_{T_0+\cdots+T_{j-1}} w, u_j), x_{j+1}\right) \\ &< \varepsilon/2 + \varepsilon/2 = \varepsilon. \end{aligned}$$

This shows that there is a controlled (ε, T)-chain from x along z to (y, w). Since by assumption $E \subset \operatorname{int} Q$ and by (3.2) this (ε, T)-chain is ε-close to an (ε, T)-chain in Q, we may choose $\varepsilon > 0$ small enough such that this is a chain in Q. This proves property (i) of chain control sets. □

The following result clarifies the relations between chain control sets and their fibers.

Proposition 3.5. *Consider system* (1.1) *in a closed subset* $Q \subset M \times \mathcal{Z}$.

(i) *Suppose that* Q *is compact, and let* $E^z \subset Q_z$, $z \in \mathcal{Z}$, *be a maximal family of sets satisfying conditions* (i)–(iii) *in Lemma* 3.2. *If* $E := \bigcup_{z \in \mathcal{Z}} E^z \subset \operatorname{int} Q$, *then* E *is a chain control set.*

(ii) *Let* E *be a chain control set. Then the fibers* E_z, $z \in \mathcal{Z}$, *are contained in a maximal family* $\tilde{E}^z \subset Q_z$, $z \in \mathcal{Z}$, *of sets satisfying conditions* (i)–(iii) *in Lemma* 3.2. *If* $\tilde{E} := \bigcup_{z \in \mathcal{Z}} \tilde{E}^z \subset \operatorname{int} Q$, *then* $E = \tilde{E}$.

Proof. It remains only to discuss the maximality properties.

(i) The union E satisfies properties (i) and (ii) of chain control sets, since for $\varepsilon < \inf_{e \in E} d(e, \partial Q)$ the controlled (ε, T)-chains are in Q. Hence, E is contained in the union $\tilde{E}$ of all sets containing E and satisfying these properties. Then $\tilde{E}$ is a chain control set and its fibers $\tilde{E}_z$ contain the sets E^z and satisfy properties (i)–(iii) in Lemma 3.2. By maximality, it follows that $E = \tilde{E}$.

(ii) Let E be a chain control set. Then the fibers E_z satisfy properties (i)–(iii) in Lemma 3.2. Clearly, the family $E_z, z \in \mathcal{Z}$, is contained in a maximal family $\tilde{E}^z, z \in \mathcal{Z}$, with these properties. If $\tilde{E} \subset \operatorname{int} Q$, the first assertion shows that $\tilde{E}$ is a chain control set and, hence, $E = \tilde{E}$. □

It is of great interest to see if the behavior in a single fiber determines chain control sets. In the periodic case, one can reconstruct chain control sets from their intersection with a fiber. More precisely, the following is a minor modification of Gayer [10], Taubert [25, Satz 2.2.5].

Proposition 3.6. *Assume that in system* (1.1) *the set* $\mathcal{Z}$ *consists of the shifts of a* T*-periodic function, and write* $Z := \mathbb{R}/T\mathbb{Z}$. *Let* $Q \subset M \times Z$ *be closed, and pick* $z_0 \in Z$. *Suppose that* $E^{z_0} \subset Q_{z_0}$ *is a maximal set such that*

(i) *for all* $x, y \in E^{z_0}$ *and all* $\varepsilon > 0$ *there are* $(x_j, z_j) \in Q \times Z$ *and controls* $u_j \in \mathcal{U}$ *with* $(x_0, z_0) = (x, z_0), (x_n, z_n) = (y, z_0)$ *such that for all* $j = 0, \ldots, n-1$

$$\mathrm{d}\left(\psi(T, (x_j, z_j, u_j)), (x_{j+1}, z_{j+1})\right) < \varepsilon \textit{ and } \psi(t, x_j, z_j, u_j) \in Q \textit{ for } t \in [0, T],$$

(ii) *for all* $x \in E^{z_0}$ *there is* $u \in \mathcal{U}$ *with* $\varphi(T, x, z_0, u), \varphi(-T, x, z_0, u) \in E^{z_0}$.

Then the set

$$E := \left\{ (x, z) \in M \times Z, \quad \begin{array}{c} \textit{there are } x_0 \in E^{z_0}, u \in \mathcal{U}, t \in [0, T) \textit{ with} \\ (x, z) = \psi(t, x_0, z_0, u) \textit{ and } \varphi(T, x_0, z_0, u) \in E^{z_0} \end{array} \right\}$$

is a chain control set relative to Q.

Conversely, for a chain control set $E \subset Q \times Z$, *every fiber* E_{z_0}, $z_0 \in Z$, *is maximal with properties* (i) *and* (ii).

In order to derive an analogous result in the almost periodic case, we have to modify property (ii) in Proposition 3.6, since it cannot be satisfied.

Theorem 3.7. *Consider system* (1.1), *and assume that* $Q \subset M \times \mathcal{Z}$ *is compact. For some* $z_0 \in \mathcal{Z}$ *let* $E^{z_0} \subset Q \times \{z_0\}$ *be a nonvoid maximal set such that for all* $x_0, y_0 \in E^{z_0}$ *and all* $\varepsilon, T > 0$ *there exists a controlled* (ε, T)*-chain in* Q *from* x_0 *along* z_0 *to* (y_0, z_0).

Then the set

$$E := \mathrm{cl}\left\{ (x,z) \in M \times Z, \begin{array}{c} \textit{for all } \varepsilon, T > 0 \textit{ there are } x_0, y_0 \in E^{z_0} \textit{ and controlled} \\ (\varepsilon, T)\textit{-chains in } Q \textit{ from } x_0 \textit{ along } z_0 \textit{ to } (y_0, z_0) \textit{ such} \\ \textit{that } (x,z) = \psi(t, x_j, \theta_{T_0+\cdots+T_{j-1}} z_0, u_j) \textit{ for some } j \\ \textit{and } t \in [0, T_j] \end{array} \right\}$$

is a chain control set relative to Q.

Proof. Consider the fibers E_z, $z \in \mathcal{Z}$, of E. By closedness of E, it is clear that $x_n \in E_{z_n}$ with $(x_n, z_n) \to (x, z) \in M \times \mathcal{Z}$ implies $x \in E_z$. Since E^{z_0} is nonvoid and E is contained in the compact set Q, hence, also compact, every fiber E_z of E is nonvoid.

Let $(x,z), (y,w) \in E$ and $\varepsilon, T > 0$. Then there exists a controlled (ε, T)-chain in Q from x along z to (y, w). This follows for elements on controlled chains from E^{z_0} to E^{z_0} by concatenating appropriate chains and using continuity (in order to guarantee $T_j \geq T$). Again, by continuity, this also follows for elements in the closure of the set of these points. It remains to show that for every $z \in \mathcal{Z}$ and every $x \in E_z$ there exists a control $u \in \mathcal{U}$ such that

$$\varphi(t, x, z, u) \in E_{\theta_t z} \text{ for all } t \in \mathbb{R}.$$

For $(x,z) \in E$ and $k \in \mathbb{N}$ choose controlled $(1/k, T)$-chains ζ^k from x along z to (x,z) with controls $u_j^k \in \mathcal{U}$. Then a subsequence of u_0^k converges to some control $v_0 \in \mathcal{U}$ and, by continuity,

$$\varphi\left(T, x, z, u_0^k\right) \to \varphi(T, x, z, v_0) \text{ for } k \to \infty.$$

Then one finds that $\varphi(T, x, z, v_0) \in E_{\theta_T z}$, since E is closed. Iterating this procedure, one constructs a control $u^+ \in \mathcal{U}$ with $\varphi(t, x, u^+) \in E$ for all $t \geq 0$. For negative times, consider the last members of the chains ζ^k. We may assume that the corresponding controls $u_{n_k}^k$ converge to a control $v \in \mathcal{U}$ and, by definition,

$$\psi\left(T_{n_k}, x_{n_k}, \theta_{T_0^k+\cdots+T_{n_k}^k} z, u_{n_k}^k\right) \to (x, z) \text{ for } k \to \infty.$$

By continuity, we may assume that $T_{n_k}^k \in [T, 2T]$ and, hence, that $T_{n_k}^k \to S \geq T$. Then $\theta_{T_{n_k}^k} u_{n_k}^k \to \theta_S v$ and continuity implies

$$\begin{aligned} &\psi\left(T_{n_k} - T, x_{n_k}, \theta_{T_0^k+\cdots+T_{n_k}^k} z, u_{n_k}^k\right) \\ &= \psi\left(-T, \psi\left(T_{n_k}, x_{n_k}, \theta_{T_0^k+\cdots+T_{n_k}^k} z, u_{n_k}^k\right), \theta_{T_{n_k}^k} u_{n_k}^k\right) \\ &\to \psi(-T, x, z, \theta_S v) \text{ for } k \to \infty. \end{aligned}$$

With $u^- := \theta_S v$ one finds that $\varphi(-T, x, z, v_1) \in E_{\theta_T z}$, since E is closed. Iterating this procedure, one constructs a control $u^- \in \mathcal{U}$ with $\varphi(t, x, z, u^-) \in E$ for all $t \leq 0$. Combining u^+ and u^- the desired control u is found. □

Remark 3.8. Theorem 3.7 shows that, up to closure, one can find chain control sets by looking at a single fiber, i.e., a single almost periodic excitation. This significantly simplifies numerical computations, since only one almost periodic excitation $z(t)$, $t \geq 0$, has to be considered. Then the resulting sets must be considered for those times T where z and $\theta_T z$ are close. In the quasi-periodic case (cf. Example 2.1), one has to look for (large) times t where all $\omega_i t$ are close to zero modulo 2π.

In addition to chain control sets E, also their projection to M defined as

$$\pi_M E := \{x \in M,\ (x,z) \in E \text{ for some } z \in \mathcal{Z}\}$$

is of interest. Obviously, for all $(x_1, x_2) \in \pi_M E$ there are $z_1, z_2 \in \mathcal{Z}$ such that $(x_1, z_1), (x_2, z_2) \in E$, and, hence, there are controlled (ε, T)-chains from x_1 along z_1 to (x_2, z_2).

4. Controllability and chain controllability. The main aim in this section is to analyze when an almost periodic solution of the uncontrolled system is contained in the interior of a subset of complete controllability. For this purpose, we ask when a reachable point is contained in the interior of the reachable set and discuss chain controllability. This leads us to control sets and their relation to chain control sets.

Again, consider control system (1.1). For a closed subset $Q \subset M \times \mathcal{Z}$, a point $x \in Q$, and $z \in \mathcal{Z}$ we define the positive and negative orbits along z relative to Q as

$$\mathcal{O}^+(x; z, Q) := \{\varphi(t, x, z, u),\ \text{with } \psi(s, x, z, u) \in Q, s \in [0, t] \text{ for some } t \geq 0, u \in \mathcal{U}\},$$
$$\mathcal{O}^-(x; z, Q) := \{\varphi(t, x, z, u),\ \text{with } \psi(s, x, z, u) \in Q, s \in [t, 0] \text{ for some } t \leq 0, u \in \mathcal{U}\}.$$

Observe that $\varphi(t, x, z, u) \in Q_{\theta_t z}$. Analogously, $\mathcal{O}_t^+(x; z, Q), \mathcal{O}_t^-(x; z, Q)$, etc. are defined, if we restrict the times accordingly. If $Q = M \times \mathcal{Z}$, we omit the argument Q.

In addition to chain control sets it is also of interest to discuss control sets, i.e., maximal subsets of approximate controllability.

Definition 4.1. *For a closed subset $Q \subset M \times \mathcal{Z}$ a subset $D \subset Q$ is a control set relative to Q if it is maximal with the following properties*:

(i) *For all $(x, z), (y, w) \in D$ there are $T_n \geq 0, u_n \in \mathcal{U}$ with $\psi(T_n, x, z, u_n) \to (y, w)$ and $\psi(t, x, z, u_n) \in Q$ for $t \in [0, T_n]$.*

(ii) *For every $z \in \mathcal{Z}$ and every $x \in D_z$ there exists a control $u \in \mathcal{U}$ such that*

$$\psi(t, x, z, u) \in D \textit{ for all } t \geq 0.$$

In condition (i), it is clear that $T_n \to \infty$, unless the excitation is periodic. Condition (ii) immediately implies that the projection of the control set is dense in $\mathcal{Z}$; the inclusion may be rewritten as $\varphi(t, x, z, u) \in D_{z(t+\cdot)}$ for all $t \geq 0$.

For periodic excitations, one can characterize control sets by looking at the discrete time system defined by the Poincaré map (Gayer [10]). We will show that also, in the almost periodic case, it is possible to characterize control sets fiberwise.

Lemma 4.2. *Suppose that $D \subset Q$ is a control set. Then the fibers $D_z := D \cap Q_z$, $z \in \mathcal{Z}$, satisfy the following properties:*

(i) *For every $z \in \mathcal{Z}$ and all $x, y \in D_z$ there are $T_n \to \infty$ and $u_n \in \mathcal{U}$ with $\psi(T_n, x, z, u_n) \to (y, z)$ and $\psi(t, x, z, u_n) \in Q$ for all $t \in [0, T_n]$.*

(ii) *For every $z \in \mathcal{Z}$ and every $x \in D_z$ there exists a control $u \in \mathcal{U}$ such that*

$$\varphi(t, x, z, u) \in D_{\theta_t z} \textit{ for all } t \geq 0.$$

Proof. The proof obviously follows from properties (i) and (ii) of control sets. ∎

The following lemma shows that the properties in Lemma 4.2 characterize control sets.

Lemma 4.3. *Suppose $Q \subset M \times \mathcal{Z}$ is closed and that $D^z \subset Q_z$, $z \in \mathcal{Z}$, is a family of sets satisfying conditions* (i) *and* (ii) *in Lemma* 4.2 *and, additionally,*

(iii) *for every $(x, z) \in D^z$ and all $T_n > 0$ with $\theta_{T_n} z \to w \in \mathcal{Z}$ there are $y \in M$ and $u_n \in \mathcal{U}$ such that $\psi(T_n, x, z, u_n) \to (y, w) \in D^w$ and $\psi(t, x, z, u_n) \in Q$ for all $t \in [0, T_n]$.*

Then $D := \bigcup_{z\in\mathcal{Z}} D^z$ *satisfies properties* (i) *and* (ii) *of control sets in Definition* 4.1.

Proof. Property (ii) of control sets is clearly satisfied due to property (ii) of the fibers. In order to prove property (i), let $(x,z),(y,w)\in D$. Since $\omega(z)=\mathcal{Z}$ there are $S_k\to\infty$ with $\theta_{S_k}z\to w$. By property (iii) we may assume that, for some controls $u_k\in\mathcal{U}$ and some $(y_0,w)\in D$,

$$\psi(S_k,x,z,u_k)\to(y_0,w)\text{ in }Q. \tag{4.1}$$

By property (i) of the fibers there are $T_n\to\infty$ and $v_n\in\mathcal{U}$ with

$$\psi(T_n,y_0,w,v_n)\to(y,w)\text{ in }Q. \tag{4.2}$$

Let $\varepsilon>0$, and denote here and in the following the open ε-ball around x by $\mathbf{B}_\varepsilon(x)$. For every $n\in\mathbb{N}$ there is an $\eta_n>0$ such that

$$\psi\left(T_n,\mathbf{B}_{\eta_n}(y_0,w),v_n\right)\subset\mathbf{B}_{\varepsilon/2}\left(\psi(T_n,y_0,w,v_n)\right) \tag{4.3}$$

due to continuous dependence on initial conditions. Convergence in (4.2) implies that $\psi(T_n,y_0,w,v_n)\in\mathbf{B}_{\varepsilon/2}(y,w)$ for sufficiently large n. Together, this yields

$$\psi\left(T_n,\mathbf{B}_{\eta_n}(y_0,w),v_n\right)\subset\mathbf{B}_\varepsilon(y,w)$$

for n large enough.

By convergence in (4.1), there is a sequence $(k_n)_{n\in\mathbb{N}}\subset\mathbb{N}$ such that

$$\psi(S_{k_n},x,z,u_{k_n})\in\mathbf{B}_{\eta_n}(y_0,w).$$

Let $\tilde{T}_n := S_{k_n}+T_n$ and

$$\tilde{u}_n(t):=\begin{cases}u_n(t) & \text{if } t<S_{k_n},\\ v_n(t-S_{k_n}) & \text{otherwise.}\end{cases}$$

Then inclusion (4.3) implies $\psi(\tilde{T}_n,x,z,\tilde{u}_n)\in\mathbf{B}_\varepsilon(y,w)$ for all $n\in\mathbb{N}$. Since $\varepsilon>0$ is arbitrary, this implies $\psi(\tilde{T}_n,x,z,\tilde{u}_n)\to(y,w)$. Furthermore, $\psi(t,x,z,\tilde{u}_n)\in Q$ for all $t\in[0,\tilde{T}_n]$, $n\in\mathbb{N}$, by construction. □

The following result clarifies the relations between control sets and their fibers.

THEOREM 4.4. *Consider system* (1.1) *in a closed subset* $Q\subset M\times\mathcal{Z}$.

(i) *Let* $D^z\subset Q_z$, $z\in\mathcal{Z}$, *be a maximal family of sets satisfying conditions* (i) *and* (ii) *in Lemma* 4.2 *and condition* (iii) *in Lemma* 4.3. *Then* $D:=\bigcup_{z\in\mathcal{Z}}D^z$ *is a control set.*

(ii) *Let* D *be a control set. Then the fibers* D_z *form a maximal family of sets satisfying conditions* (i) *and* (ii) *in Lemma* 4.2.

Proof. By Lemmas 4.2 and 4.3 only maximality has to be shown.

(i) By Lemma 4.3, the set $D:=\bigcup_{z\in\mathcal{Z}}D^z$ satisfies the two defining properties of control sets and is, thus, contained in a control set $\tilde{D}$. The fibers $\tilde{D}_z$, $z\in\mathcal{Z}$, satisfy conditions (i) and (ii) in Lemma 4.2. So, by maximality, $\tilde{D}_z=D^z$ for every $z\in\mathcal{Z}$, which implies $D=\tilde{D}$.

(ii) By Lemma 4.2 the fibers D_z satisfy conditions (i) and (ii) and are, thus, contained in a maximal family D^z, $z\in\mathcal{Z}$, of sets satisfying these properties. By Lemma 4.3 the set $\tilde{D}:=\bigcup_{z\in\mathcal{Z}}D^z$ is a control set. Clearly $D\subset\tilde{D}$. Maximality implies $D=\tilde{D}$, and so $D_z=D^z$ for all z. □

We note the following simple property of control sets.

PROPOSITION 4.5. *Let D_1 and D_2 be control sets relative to Q, and assume that there are $z \in \mathcal{Z}$, times $T_2 > T_1 > 0$, a point $x \in D_1^z$, and a control $u \in \mathcal{U}$ such that*

$$\begin{aligned} &\varphi(T_1, x, z, u) \in D_{2,z(T_1+\cdot)} \text{ and } \varphi(T_1 + T_2, x, z, u) \in D_{1,z(T_1+T_2+\cdot)}, \\ &\text{and } \psi(t, x_1, z, u) \in Q \text{ for all } t \in [0, T_1 + T_2]. \end{aligned}$$

Then $D_1 = D_2$.

Proof. The proof follows by maximality of D_1, since $D_1 \cup \{\psi(t, x, z, u), t \in [0, T_1 + T_2]\}$ satisfies properties (i) and (ii) of control sets. □

Our next aim is to prove that under an inner-pair condition every almost periodic solution of the uncontrolled equation is contained in the interior of a control set. For a periodic excitation as considered in Example 2.1, the state space $Z = \mathbb{S}^1$ is (trivially) completely controllable. However, already for a quasi-periodic excitation with two noncommensurable (i.e., rationally independent) frequencies ω_1, ω_2, this is no longer true. Hence, it does not make sense to consider exact controllability properties in the z-component. This is different in the x-component as shown by the following proposition.

PROPOSITION 4.6. *Let $\psi(t, x^0, z^0, 0) \in Q$, $t \in \mathbb{R}$, be an almost periodic solution of the uncontrolled system, and define $A := \mathrm{cl}\{\psi(t, x^0, z^0, 0), t \in \mathbb{R}\}$. Assume that there are $\varepsilon, T > 0$ such that for every $(x, z) \in A$*

$$\mathbf{B}_\varepsilon\left(\varphi(T, x, z, 0)\right) \subset \mathcal{O}_T^+(x; z, Q).$$

Then for all $(x, z), (y, w) \in A$ there is $\tau > 0$ such that $\mathbf{B}_{\varepsilon/2}(y) \subset \mathcal{O}_\tau^+(x; z, Q)$, and for every $y_0 \in \mathbf{B}_{\varepsilon/2}(y)$ there are $\tau_n \geq 0$ and $u_n \in \mathcal{U}$ with $\varphi(\tau_n, x, z, u_n) = y_0$ in Q and $\theta_{\tau_n} z \to w$.

Proof. Let $(x, z), (y, w) \in A$. Note that, by uniform continuity, there is $\delta > 0$ such that

$$\mathrm{d}\left((x_1, z_1), (x_2, z_2)\right) < \delta \text{ implies } \mathrm{d}\left(\psi(T, x_1, z_1, 0), \psi(T, x_2, z_2, 0)\right) < \varepsilon/2.$$

By almost periodicity, one has $\omega(x, z) = A$, and, hence, there are $S_n \to \infty$ such that $\psi(S_n, x, z, 0) \to \psi(-T, y, w, 0)$ in $A \subset Q$. Choose n large enough such that for $S_0 := S_n$

$$\mathrm{d}(\psi(-T, y, w, 0), \psi(S_0, x, z, 0)) < \delta. \tag{4.4}$$

This implies

$$\mathrm{d}((y, w), \psi(S_0 + T, x, z, 0)) = \mathrm{d}(\psi(T, \psi(-T, y, w, 0), 0), \psi(T, \psi(S_0, x, z, 0), 0)) < \varepsilon/2,$$

and we conclude for $\varepsilon > 0$, small enough,

$$\begin{aligned} \mathbf{B}_{\varepsilon/2}(y) &\subset \mathbf{B}_\varepsilon\left(\varphi(S_0 + T, x, z, 0)\right) = \mathbf{B}_\varepsilon\left(\varphi\left(T, \varphi(S_0, x, z, 0), \theta_T z\right)\right) \\ &\subset \mathrm{int}\, \mathcal{O}_T^+\left(\varphi(S_0, x, z, 0); \theta_T z, Q\right) \subset \mathrm{int}\, \mathcal{O}_{S_0+T}^+(x; z, Q). \end{aligned}$$

This yields the first assertion with $\tau = S_0 + T$, and the second assertion follows with $\tau_n := S_n + T$ if we consider $\delta_n \to 0$ in (4.4). □

This proposition allows us to show that almost periodic solutions of the uncontrolled system are contained in the interior of control sets. In other words, around an almost periodic solution we have complete controllability along the almost periodic excitations.

THEOREM 4.7. *Let $\psi(t,x^0,z^0,0) \in Q$, $t \in \mathbb{R}$, be an almost periodic solution of the uncontrolled system, and let $A := \mathrm{cl}\{\psi(t,x^0,z^0,0), t \in \mathbb{R}\}$. Assume that there are $\varepsilon, T > 0$ such that for every $(x,z) \in A$*

$$\mathbf{B}_\varepsilon(\varphi(T,x,z,0)) \subset \mathcal{O}_T^+(x;z,Q) \text{ and } \mathbf{B}_\varepsilon(\varphi(-T,x,z,0)) \subset \mathcal{O}_T^-(x;z,Q). \tag{4.5}$$

Then there exists a control set D such that for every $(x,z) \in A$ one has $x \in \operatorname{int} D^z$.

Proof. It is clear that the set A satisfies properties (i) and (ii) of Definition 4.1. Hence, it is contained in a maximal set with these properties, i.e., a control set D. The assertion follows if we can show that for all $(x,z) \in A$ the neighborhoods $\mathbf{B}_{\varepsilon/2}(x)$ also satisfy these properties. Let $(x,z),(y,w) \in A$. For property (i) it suffices to show that for $x_0 \in \mathbf{B}_{\varepsilon/2}(x)$, $y_0 \in \mathbf{B}_{\varepsilon/2}(y)$ there are $T_n \geq 0$ and $u_n \in \mathcal{U}$ with $\psi(T_n,y_0,w,u_n) \to (x_0,z)$ in Q. Since $\psi(T,x,z,0) \in A$, condition (4.5) implies

$$\mathbf{B}_{\varepsilon/2}(x) \subset \mathcal{O}_T^-\left(\psi(T,x,z,0)\right).$$

Hence, for every $(x_0,z) \in \mathbf{B}_{\varepsilon/2}(x) \times \{z\}$ there is a control $u_0 \in \mathcal{U}$ with $\psi(T,x,z,0) = \psi(T,x_0,z,u_0)$. Similarly, $\psi(-T,y,w,0) \in A$ implies

$$\mathbf{B}_{\varepsilon/2}(y) \subset \mathcal{O}_T^+\left(\psi(-T,y,w,0)\right),$$

and, hence, there is a control $v_0 \in \mathcal{U}$ with $(y_0,w) = \psi(T,\psi(-T,y,w,0),v_0)$.

Since $\psi(T,x,z,0), \psi(-T,y,w,0) \in A$ there are $S_n \geq 0$ and $v_n \in \mathcal{U}$ with

$$\psi\left(S_n, \psi(T,x,z,0), v_n\right) \to \psi(-T,y,w,0) \text{ in } Q.$$

By continuity, this implies

$$\psi\left(T, \psi\left(S_n, \psi(T,x,z,0), v_n\right), v_0\right) \to \psi\left(T, \psi(-T,y,w,0), v_0\right) = (y_0,w).$$

Define the concatenated controls

$$u_n(t) := \begin{cases} u_0(t) & \text{for} \quad t \in [0,T], \\ v_n(t-T) & \text{for} \quad t \in (T, T+S_n], \\ v_0(t-T-S_n) & \text{for} \quad t \in (T+S_n, 2T+S_n]. \end{cases}$$

Then, with $T_n := 2T + S_n$,

$$\begin{aligned} \psi(T_n,x_0,z,u_n) &= \psi(2T+S_n,x_0,z,u_n) \\ &= \psi\left(T, \psi\left(S_n, \psi(T,x_0,z,u_0), v_n\right), v_0\right) \\ &= \psi\left(T, \psi\left(S_n, \psi(T,x,z,0), v_n\right), v_0\right) \\ &\to (y_0,w). \end{aligned}$$

This proves property (i). Then property (ii) is obvious. □

Remark 4.8. Condition (4.5) is analogous to the inner-pair condition (but slightly stronger) for autonomous control systems; see [4, Definition 4.1.5].

Assumption (4.5) in Theorem 4.7 can be guaranteed for a large class of systems, as shown by Gayer [10]: Consider the following nth order systems on $\mathbb{R}^m$:

$$\begin{pmatrix} x_1^{(n)} \\ \vdots \\ x_m^{(n)} \end{pmatrix} + \begin{pmatrix} f_1\left(t,x,\dots,x^{(n-1)}\right) \\ \vdots \\ f_m\left(t,x,\dots,x^{(n-1)}\right) \end{pmatrix} = \begin{pmatrix} b_1\left(t,x,\dots,x^{(n-1)}\right) & u_1(t) \\ & \vdots \\ b_m\left(t,x,\dots,x^{(n-1)}\right) & u_m(t) \end{pmatrix}. \tag{4.6}$$

Here $x = (x_i) \in C^{n-1}(\mathbb{R}, \mathbb{R}^m)$, its nth derivative exists but is not necessarily continuous, and $x^{(k)}$ denotes its kth derivative. Assume $f_i : \mathbb{R} \times \mathbb{R}^{nm} \to \mathbb{R}$ and $b_i : \mathbb{R} \times \mathbb{R}^{nm} \to \mathbb{R}$ are C^1, and consider controls

$$u = (u_i) \in \mathcal{U}^\rho := \{u : \mathbb{R} \to \mathbb{R}^m,\ u(t) \in U^\rho \text{ for all } t\}.$$

We assume that the control ranges U^ρ are compact and convex and that mapping $\rho \mapsto U^\rho$ is strictly increasing, i.e., $U^{\rho_1} \subset \operatorname{int} U^{\rho_2}$ for $0 \le \rho_1 \le \rho_2$. As before, assume that for all initial values and all controls the solutions are unique and exist for all times.

We consider the associated first order systems. So for initial values $y_0, \dots, y_{n-1} \in \mathbb{R}^m$ at time $t_0 = 0$ and a control $u \in \mathcal{U}^\rho$, denote by $\lambda(t, y_0, \dots, y_{n-1}, u)$ the corresponding solution of (4.6). We set $y^0 := (y_0, \dots, y_{n-1}) \in \mathbb{R}^{nm}$ and define the set reachable from y^0 at time $T > 0$ by

$$\mathcal{O}_T^{+,\rho}(y^0) := \left\{ \begin{matrix} (z_0, \dots, z_{n-1}) \in \mathbb{R}^{nm}, \text{ there is } u \in \mathcal{U} \\ \text{with } z_i = \lambda^{(i)}(t, y^0) \text{ for } 0 \le i \le n-1 \end{matrix} \right\}.$$

Proposition 4.9. *Consider system* (4.6)*, and assume that there is some $\alpha > 0$ such that $|b_i(t, y)| \ge \alpha$ for all $i \in \{1, \dots, m\}$ and all $(t, y) \in \mathbb{R} \times \mathbb{R}^{nm}$. Let $0 \le \rho_1 \le \rho_2$, and consider a compact subset $B \subset \mathbb{R}^{nm}$. Then for every $T > 0$ there is $\varepsilon > 0$ such that for all $(y^0, u) \in B \times \mathcal{U}^{\rho_1}$*

$$\mathbf{B}\left(\left(\lambda\left(T, y^0, u\right), \dots, \lambda^{(n-1)}\left(T, y^0, u\right)\right); \varepsilon\right) \subset \mathcal{O}_T^{+,\rho_2}\left(y^0\right).$$

Proof. The proof follows from [10, Theorem 3] and its proof. Here arbitrary time dependence of the right-hand side is allowed and the theorem is formulated a bit differently (in terms of inner pairs for varying control range), but the proof shows the stronger result formulated above. □

In particular, under the assumptions of Proposition 4.9, one obtains for $\rho_1 = 0$ that condition (4.5) is satisfied (applying the theorem also to the time reversed system).

Next, we generalize Theorem 4.7 in order to show a relation between chain controllability and controllability. We begin with the following lemma.

Lemma 4.10. *Let $0 \le \rho_1 \le \rho_2$, and consider a compact subset $Q \subset M \times \mathcal{Z}$. Let E^{ρ_1} be a chain control set relative to Q for system* (1.1) *with controls in $\mathcal{U}^{\rho_1}$. Assume that there are $\varepsilon, T > 0$ such that for every $(x, z) \in E^{\rho_1}$ and $u \in \mathcal{U}^{\rho_1}$*

$$\mathbf{B}_\varepsilon(\varphi(T, x, z, u)) \subset \mathcal{O}_T^{+,\rho_2}(x; z, Q). \tag{4.7}$$

Then for all $(x, z), (y, w) \in E^{\rho_1}$ there is $\tau > 0$ such that $\mathbf{B}_{\varepsilon/2}(y) \subset \mathcal{O}_\tau^{+,\rho_2}(x; z, Q)$, and for every $y_0 \in \mathbf{B}_{\varepsilon/2}(y)$ there are $\tau_n \ge 0$ and $u_n \in \mathcal{U}^{\rho_2}$ with $\varphi(\tau_n, x, z, u_n) = y_0$ in Q and $\theta_{\tau_n} u_n \to w$.

Proof. Let $(x, z), (y, w) \in E^{\rho_1}$. By uniform continuity, there is δ with $0 < \delta < \varepsilon/2$ such that for all u

$$\mathrm{d}\left((x_1, z_1), (x_2, z_2)\right) < \delta \text{ implies } \mathrm{d}\left(\psi(T, x_1, z_1, u), \psi(T, x_2, z_2, u)\right) < \varepsilon/2.$$

There is $u_0 \in \mathcal{U}^{\rho_1}$ such that $\psi(-T, y, w, u_0) \in E^{\rho_1}$. By chain controllability, there exists a controlled (δ, T)-chain in Q along z from x to $\psi(-T, y, w, u_0)$, i.e., $x_0 = x$, $x_n = \varphi(-T, y, w, u_0)$, and

$$\mathrm{d}(\theta_{T_0 + \cdots + T_{n-1}} z, \theta_{-T} w) < \delta,\ \mathrm{d}\left(\varphi(T_j, x_j, \theta_{T_0 + \cdots + T_{j-1}} z, v_j), x_{j+1}\right) < \delta \text{ for all } j,$$
$$\psi\left(t, x_j, \theta_{T_0 + \cdots + T_{j-1}} z, v_j\right) \in Q \text{ for all } t \in [0, T_j] \text{ and for all } j.$$

For every j, one finds by induction

$$\begin{aligned} x_{j+1} &\in \mathbf{B}_\delta\left(\varphi\left(T_j, x_j, \theta_{T_0+\cdots+T_{j-1}}z, v_j\right)\right) \\ &= \mathbf{B}_\delta\left(\varphi\left(T, \varphi\left(T_j - T, x_j, \theta_{T_0+\cdots+T_{j-1}}z, v_j\right), \theta_{T_0+\cdots+T_{j-1}+T_j-T}z, \theta_{T_j-T}v_j\right)\right) \\ &\subset \mathcal{O}_T^{+,\rho_2}\left(\varphi\left(T_j - T, x_j, \theta_{T_0+\cdots+T_{j-1}}z, v_j\right); \theta_{T_0+\cdots+T_{j-1}+T_j-T}z, Q\right) \\ &\subset \mathcal{O}^{+,\rho_2}_{T_0+\cdots+T_j}(x_0; z, Q). \end{aligned}$$

Hence, there is a control $v \in \mathcal{U}^{\rho_2}$ with

$$x_n = \varphi(T_0 + \cdots + T_{n-1}, x, z, v) \text{ and } \mathrm{d}(\theta_{T_0+\cdots+T_{n-1}}z, \theta_{-T}w) < \delta. \tag{4.8}$$

By choice of δ we find

$$\begin{aligned} &\mathrm{d}\left(\psi\left(T, x_n, \theta_{T_0+\cdots+T_{n-1}}z, \theta_{-T}u_0\right), (y, w)\right) \\ &= \mathrm{d}\left(\psi\left(T, x_n, \theta_{T_0+\cdots+T_{n-1}}z, \theta_{-T}u_0\right), \psi\left(T, \psi(-T, y, w, u_0), \theta_{-T}u_0\right)\right) < \varepsilon/2. \end{aligned}$$

We conclude for $\varepsilon > 0$, small enough,

$$\begin{aligned} \mathbf{B}_{\varepsilon/2}(y) &\subset \mathbf{B}_\varepsilon\left(\varphi\left(T, x_n, \theta_{T_0+\cdots+T_{n-1}}z, \theta_{-T}u_0\right)\right) \\ &= \mathbf{B}_\varepsilon\left(\varphi\left(T, \varphi\left(T_0 + \cdots + T_{n-1}, x, z, v\right), \theta_{T_0+\cdots+T_{n-1}}z, \theta_{-T}u_0\right)\right) \\ &\subset \mathcal{O}^{+,\rho_2}_{T_0+\cdots+T_{n-1}+T}(x; z, Q). \end{aligned}$$

This yields the first assertion with $\tau = T_0 + \cdots + T_{n-1} + T$. The second assertion follows with $\tau_n = T_0 + \cdots + T_{n-1} + T$ if we consider $\delta_n \to 0$ in (4.8). □

This lemma allows us to show that chain control sets are contained in the interior of control sets for larger control ranges.

THEOREM 4.11. *Let $0 \le \rho_1 \le \rho_2$, and consider a compact subset $Q \subset M \times \mathcal{Z}$. Let E^{ρ_1} be a chain control set relative to Q for system* (1.1) *with controls in $\mathcal{U}^{\rho_1}$. Assume that there are $\varepsilon, T > 0$ such that for every $(x, z) \in E^{\rho_1}$ and $u \in \mathcal{U}^{\rho_1}$*

$$\mathbf{B}_\varepsilon\left(\varphi(T, x, z, u)\right) \subset \mathcal{O}_T^{+,\rho_2}(x; z, Q) \text{ and } \mathbf{B}_\varepsilon\left(\varphi(-T, x, z, u)\right) \subset \mathcal{O}_T^{-,\rho_2}(x; z, Q). \tag{4.9}$$

Then there exists a control set D^{ρ_2} such that for every $(x, z) \in E^{\rho_1}$ one has $x \in \operatorname{int} D_z^{\rho_2}$.

Proof. The assertion follows if we can show that for all $(x, z) \in E^{\rho_1}$ the neighborhoods $\mathbf{B}_{\varepsilon/2}(x)$ satisfy conditions (i) and (ii) in Definition 4.1 for controls in $\mathcal{U}^{\rho_2}$. Then E^{ρ_1} is contained in a maximal set with these properties, i.e., a control set D^{ρ_2}. Let $(x, z), (y, w) \in E^{\rho_1}$. For property (i) it suffices to show that for $x_0 \in \mathbf{B}_{\varepsilon/2}(x)$, $y_0 \in \mathbf{B}_{\varepsilon/2}(y)$ there are $T_n \ge 0$ and $u_n \in \mathcal{U}^{\rho_2}$ with $\psi(T_n, y_0, w, u_n) \to (x_0, z)$ in Q. There is a control $v_0 \in \mathcal{U}^{\rho_1}$ with $\psi(T, x, z, v_0) \in E^{\rho_1}$, and, hence, condition (4.5) implies

$$\mathbf{B}_{\varepsilon/2}(x) \subset \mathcal{O}_T^{-,\rho_2}\left(\psi(T, x, z, v_0)\right).$$

Hence, for every $x_0 \in \mathbf{B}(x, \varepsilon/2)$ there is a control $u_0 \in \mathcal{U}^{\rho_2}$ with $\psi(T, x, z, v_0) = \psi(T, x_0, z, u_0)$. Similarly, there is a control $v_1 \in \mathcal{U}^{\rho_1}$ with $\psi(-T, y, w, v_1) \in E^{\rho_1}$ and

$$\mathbf{B}_{\varepsilon/2}(y) \subset \mathcal{O}_T^{+,\rho_2}\left(\psi(-T, y, w, v_1)\right),$$

and, hence, there is a control $u_1 \in \mathcal{U}^{\rho_2}$ with $(y_0, w) = \psi(T, \psi(-T, y, w, v_1), u_1)$.

Since $\psi(T, x, z, v_0), \psi(-T, y, w, v_1) \in E^{\rho_1}$, Lemma 4.10 implies that there are $\tau_n \ge 0$ and $v_n \in \mathcal{U}^{\rho_2}$ with $\psi(\tau_n, \psi(T, x, z, v_0), v_n) \to \psi(-T, y, w, v_1)$ in Q.

Together, one obtains

$$\psi\left(T, \psi\left(\tau_n, \psi(T, x, z, v_0), v_n\right), u_1\right) \to \psi\left(T, \psi(-T, y, w, v_1), u_1\right) = (y_0, w).$$

Define the concatenated control $u_n \in \mathcal{U}^{\rho_2}$ by

$$u_n(t) := \begin{cases} u_0(t) & \text{for} \quad t \in [0, T], \\ v_n(t - T) & \text{for} \quad t \in (T, T + \tau_n], \\ u_1(t - T - \tau_n) & \text{for} \quad t \in [T + \tau_n, 2T + \tau_n]. \end{cases}$$

Then, with $T_n := 2T + \tau_n$,

$$\begin{aligned} \psi(T_n, x_0, z, u_n) &= \psi(2T + \tau_n, x_0, z, u_n) \\ &= \psi\left(T, \psi\left(\tau_n, \psi(T, x_0, z, u_n), \theta_T u_n\right), \theta_{T+\tau_n} u_n\right) \\ &= \psi\left(T, \psi\left(\tau_n, \psi(T, x_0, z, u_0), v_n\right), u_1\right) \\ &\to (y_0, w). \end{aligned}$$

This proves property (i) of control sets. Now property (ii) is obvious. □

Remark 4.12. Using this theorem we can, as in [4, Theorem 4.7.5], show that for all up to at most countably many ρ-values the closures of control sets and the chain control sets coincide. The proof is based on Scherbina's lemma [20] for continuity of monotonically increasing set valued functions. Hence, by Theorem 3.7 one may also determine the fibers of control sets via the fibers of the chain control sets. For this purpose, one has to consider "long" times, since these fibers are determined only on long time intervals; cf. Remark 3.8. At first sight, this is different if the excitation is periodic; here only the Poincaré map and, hence, the period length are needed; see Proposition 3.6. Nevertheless, also in this case approximate controllability is relevant (the entrance boundary of a control set is reached from the interior only for time tending to infinity), and, hence, also these objects are determined only on long time intervals.

5. Almost periodic solutions and heteroclinic orbits.

In this section, we recall results on almost periodic perturbations of hyperbolic equilibria and Melnikov's method. Since in the literature they are not precisely stated in the form needed here, we recall the relevant concepts and some arguments for the proofs.

It is well-known that, under small periodic perturbations, a hyperbolic fixed point of an autonomous differential equation becomes a periodic solution; see, e.g., [1, Theorem 25.2] for details on this result, which is known as the *Poincaré continuation.* This result can be generalized to almost periodic perturbations, in which case the existence of an almost periodic solution can be shown. Consider the differential equation

$$\dot{x} = g(x) + \mu h(t, x, \mu) \tag{5.1}$$

for $g : \mathbb{R}^d \to \mathbb{R}^d$ and $h : \mathbb{R} \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$. The parameter $\mu \in \mathbb{R}$ is interpreted as a small perturbation. Setting $\mu = 0$ in system (5.1) leads to the equation $\dot{x} = g(x)$, which will be referred to as the *unperturbed* system. Throughout, we assume that (5.1) satisfies the following conditions: The function g is C^1, h is continuous, h_x exists, and there are a bounded and open subset $V \subset \mathbb{R}^d$ containing x_0 and a constant $\bar{\mu} > 0$ such that h and h_x are almost periodic in t, uniformly with respect to $(x, \mu) \in \operatorname{cl} V \times [-\bar{\mu}, \bar{\mu}]$, and solutions of (5.1) exist for all starting points in V, all $\mu \in [-\bar{\mu}, \bar{\mu}]$, and all times.

As noted in Scheurle [21, Remark 2.7], almost periodicity of h_x uniformly with respect to (x, μ) is equivalent with h_x being uniformly continuous on $\mathbb{R} \times \operatorname{cl} V \times [-\bar{\mu}, \bar{\mu}]$.

Next, recall the notion of exponential dichotomies, which generalize the idea of hyperbolicity to nonautonomous systems; cf. Coppel [6].

DEFINITION 5.1. *Consider the system*

$$\dot{x} = A(t)x \tag{5.2}$$

for a piecewise continuous matrix function $A : J \to \mathbb{R}^{d\times d}$ *defined on an interval* $J \subset \mathbb{R}$*, and let* $X(t)$ *be a fundamental matrix function for* (5.2). *System* (5.2) *has an* exponential dichotomy *on* J *if there is a projection* $P : \mathbb{R}^d \to \mathbb{R}^d$ *and constants* $K \geq 1$, $\alpha > 0$ *such that*

$$\begin{aligned} \left\|X(t)PX^{-1}(s)\right\| &\leq Ke^{-\alpha(t-s)} && \textit{for } s \leq t, \\ \left\|X(t)(I-P)X^{-1}(s)\right\| &\leq Ke^{-\alpha(s-t)} && \textit{for } s \geq t. \end{aligned}$$

Then the following perturbation result [21, Lemma 2.4] holds.

LEMMA 5.2. *Let* $g(t,x)$ *and* $h(t,x)$ *be functions which are defined and continuous on* $\mathbb{R}\times V$ *with values in* $\mathbb{R}^d$*, where* V *is an open subset of* $\mathbb{R}^d$*. Furthermore, assume that the partial derivatives* g_x *and* h_x *exist,* g_x *is uniformly continuous, and* h_x *is continuous in* $\mathbb{R}\times V$*. Finally, assume that the equation* $\dot{x} = g(t,x)$ *has a solution* $x = x_0(t)$ *defined and contained in* V *for all* $t \in \mathbb{R}$ *and strictly bounded away from the boundary of* V *such that the variational equation* $\dot{x} = g_x(t, x_0(t))x$ *has an exponential dichotomy on* $\mathbb{R}$ *with constants* K *and* α*. Then there exist a positive constant* η_0 *and a function* $\eta_1(\eta)$ *depending only on* g*,* K*, and* α *such that if* $0 < \eta \leq \eta_0$*,*

$$\sup_{(t,x)\in\mathbb{R}\times V} \|h(t,x)\| < \eta_1(\eta) \textit{ and } \sup_{(t,x)\in\mathbb{R}\times V} \|h_x(t,x)\| < K\ \alpha/2,$$

then the equation $\dot{x} = g(t,x) + h(t,x)$ *has a unique solution* $x(t)$ *satisfying* $\|x(t) - x_0(t)\| \leq \eta$, $t \in \mathbb{R}$.

A slight modification of Bohr's proof for the boundedness of almost periodic functions in [2] shows uniform boundedness of uniformly almost periodic functions.

LEMMA 5.3. *Let* Λ *be a compact topological space,* M *a normed vector space with norm* $\|\cdot\|$*, and* $f : \mathbb{R}\times\Lambda \to M$ *continuous and almost periodic in* t *uniformly with respect to* $x \in \Lambda$*. Then*

$$\sup{}_{(t,x)\in\mathbb{R}\times\Lambda}\|f(t,x)\| < \infty.$$

Proof. Since f is uniformly almost periodic, there is an interval length L such that for every interval $J \subset \mathbb{R}$ of length L there exists a translation number $\tau(J) \in J$ satisfying $\|f(t+\tau(J),x) - f(t,x)\| < 1$ for all $(t,x) \in \mathbb{R}\times\Lambda$. Here L and τ are independent of x due to uniformity.

Since f is continuous and Λ compact, $c := \sup_{(t,x)\in[0,L]\times\Lambda}\|f(t,x)\| < \infty$. For every $t \in \mathbb{R}$ any translation number τ_t in the interval $J = [-t, -t+L]$ satisfies $t + \tau_t \in [0, L]$. Therefore, for every $t \in \mathbb{R}$ and $x \in \Lambda$

$$\|f(t,x)\| \leq \|f(t+\tau_t)\| + \|f(t) - f(t+\tau_t)\| \leq c+1. \qquad \square$$

The previous lemmas imply the following result (this is essentially Lemma 2.8 in [21]).

PROPOSITION 5.4. *Suppose that the unperturbed system corresponding to* (5.1) *has a hyperbolic fixed point* x_0*; i.e.,* $g(x_0) = 0$ *and the real parts of the eigenvalues of* $g_x(x_0)$ *are different from* 0*. For all (small)* $\eta > 0$ *there is* $\mu_0 = \mu_0(\eta) > 0$ *such that for*

$|\mu| \leq \mu_0$ there exists a unique solution $\zeta^\mu(t)$ of system (5.1) *satisfying $\|\zeta^\mu(t)-x_0\| \leq \eta$ for all $t \in \mathbb{R}$. This solution is almost periodic.*

Proof. First, we show that system (5.1) satisfies the assumptions of Lemma 5.2. The functions g and h are continuous, the derivatives g_x and h_x exist, and g_x is uniformly continuous on the compact set $\operatorname{cl} V$. As x_0 is a hyperbolic equilibrium of the unperturbed equation, the corresponding linearized equation $\dot{x} = g_x(x_0)x$ trivially has an exponential dichotomy on $\mathbb{R}$. Finally, $\sup_{(t,x)\in\mathbb{R}\times V} \|\mu h(t,x,\mu)\|$ and $\sup_{(t,x)\in\mathbb{R}\times V} \|\mu h_x(t,x,\mu)\|$ can be made arbitrarily small by choosing μ small enough, since h and h_x are uniformly almost periodic and, thus, uniformly bounded, due to Lemma 5.3.

This means that for sufficiently small perturbations μ there is a unique solution ζ^μ which stays near the original fixed point x_0 for all times. For sufficiently small μ the equation

$$\dot{x} = \left[g_x\left(\zeta^\mu(t)\right) + \mu h_x\left(t, \zeta^\mu(t), \mu\right)\right] x$$

has an exponential dichotomy on $\mathbb{R}$. This follows from roughness of exponential dichotomies with respect to small perturbations; see [21, Proposition 2.2] or [6, p. 34]. Finally, it remains to show almost periodicity of the perturbed solution ζ^μ. For this purpose, consider the shifted system

$$\dot{x} = g(x) + \mu h(t+\tau, x, \mu) \tag{5.3}$$

for $\tau \in \mathbb{R}$. Lemma 5.2 applied to (5.3) shows that for small η and $|\mu| \leq \mu_0(\eta)$ there is a unique solution $\zeta^\mu_\tau(t)$ which satisfies $\|\zeta^\mu_\tau(t) - x_0\| \leq \eta$ for all $t \in \mathbb{R}$. Obviously, $\zeta^\mu_\tau(t) = \zeta^\mu(t+\tau)$ for all $t, \tau \in \mathbb{R}$.

Now we apply Lemma 5.2 to (5.3) again, setting $g(t,x) = g(x) + \mu h(t,x,\mu)$, $h(t,x) = \mu[h(t+\tau,x,\mu) - h(t,x,\mu)]$, and $x_0(t) = \zeta^\mu(t)$. For sufficiently small μ and $\eta > 0$, there is an $\varepsilon = \varepsilon(\mu,\eta) > 0$ such that $\|\zeta^\mu(t) - \zeta^\mu_\tau(t)\| \leq \eta$, provided that

$$|\mu| \sup_{(t,x)\in\mathbb{R}\times V} \|h(t+\tau,x,\mu) - h(t,x,\mu)\| < \varepsilon$$

and

$$|\mu| \sup_{(t,x)\in\mathbb{R}\times V} \|h_x(t,x,\mu) - h_x(t+\tau,x,\mu)\| < \varepsilon.$$

Hence, uniform almost periodicity of h and h_x implies almost periodicity of $\zeta^\mu(t)$. □

If we suppose that in our setting there exist two hyperbolic fixed points $x_\pm \in \mathbb{R}^d$ of the unperturbed system, Proposition 5.4 implies the existence of almost periodic solutions $\zeta^\mu_\pm$ near $x_\pm$ for sufficiently small μ. If there is a heteroclinic orbit ζ from x_- to x_+, the question arises how the system behaves near ζ for small perturbations μ.

For time-periodic perturbations Melnikov's method gives a handy criterion for the existence of transversal heteroclinic points. Palmer has developed a generalization of Melnikov's method in [18] which, in our setting, yields the following theorem.

Theorem 5.5. *Consider the system $\dot{x} = g(x) + \mu h(t,x,\mu)$, and let the following assumptions be satisfied:*

(i) *There are a bounded and open subset $V \subset \mathbb{R}^d$ and a constant $\bar{\mu} > 0$ such that $g : V \to \mathbb{R}^d$ is C^2 and $h : \mathbb{R} \times V \times [-\bar{\mu}, \bar{\mu}] \to \mathbb{R}^d$ is continuous. The partial derivatives h_t, h_x, h_μ, h_{xx}, $h_{x\mu}$, $h_{\mu x}$, and $h_{\mu\mu}$ exist and are bounded, continuous in t for each fixed x, μ, and continuous in x, μ uniformly with respect to t, x, and μ.*

(ii) *The functions h and h_x are almost periodic in t, uniformly with respect to $(x,\mu) \in \operatorname{cl} V \times [-\bar{\mu},\bar{\mu}]$.*

(iii) *The unperturbed equation $\dot{x} = g(x)$ has hyperbolic fixed points $x_\pm \in V$ with stable and unstable manifolds of the same dimensions.*

(iv) *There is a heteroclinic orbit ζ from x_- to x_+ contained in V.*

(v) *The function*

$$\Delta(t_0) := \int_{-\infty}^{\infty} \varphi(t) \cdot h\left(t + t_0, \zeta(t), 0\right)\ \mathrm{d}t$$

has a simple zero at some $t_0 \in \mathbb{R}$, where $\varphi(t)$ is the unique (up to a scalar multiple) bounded solution of the adjoint system $\dot{x} = g_x(\zeta(t))^T x$ and "$\cdot$" denotes the inner product in $\mathbb{R}^d$.

Then there exists $\delta_0 > 0$ such that for sufficiently small μ the perturbed system (5.1) has a unique solution $x(t,\mu)$ satisfying $\|x(t,\mu) - \zeta(t-t_0)\| \leq \delta_0$ for all $t \in \mathbb{R}$. Furthermore,

$$\sup_{t\in\mathbb{R}} \|x(t,\mu) - \zeta(t-t_0)\| = O(\mu) \text{ for } \mu \to 0$$

holds, and

$$\dot{x} = \left[g_x\left(x(t,\mu)\right) + \mu h_x\left(t, x(t,\mu), \mu\right)\right] x$$

has an exponential dichotomy on $\mathbb{R}$.

Finally, it holds that

$$\lim_{t\to\pm\infty} \|x(t,\mu) - \zeta_\pm^\mu(t)\| = 0 \tag{5.4}$$

for sufficiently small μ, where $\zeta_\pm^\mu$ are the almost periodic solutions near $x_\pm$.

Proof. The proof follows from [18, Corollary 4.3] and the remark on pp. 251–252 in [18] combined with the ideas of the proof of [18, Corollary 4.4] using the fact that $\dot{x} = g_x(\zeta(t))x$ has an exponential dichotomy on both half-lines and that the dimensions of the stable and unstable subspaces sum up to d.

More precisely, [18, Corollary 4.4] shows (5.4) for the *periodic* case. But, in fact, periodicity is needed only there to prove periodicity of $\zeta_\pm^\mu$. So (5.4) holds for the almost periodic case, too; cf. [21, Remark 2.9]. In detail, there is a $\delta > 0$ independent of μ such that if

$$\|x(t,\mu) - \zeta_\pm^\mu(t)\| \leq \delta \tag{5.5}$$

for sufficiently large $|t|$ (positive for "+", negative for "−"), then (5.4) holds; cf. [11, Theorem 3.1]. For sufficiently small μ and large $|t|$

$$\left\|x(t,\mu) - \zeta_\pm^\mu(t)\right\| \leq \|x(t,\mu) - \zeta(t-t_0)\| + \|\zeta(t-t_0) - x_\pm\| + \left\|x_\pm - \zeta_\pm^\mu(t)\right\| \leq \delta,$$

and, hence, (5.5) holds.

The fact that the variational system $\dot{x} = g_x(\zeta(t))x$ has an exponential dichotomy and that the dimensions sum up to d follows from standard perturbation theory and from the assumption that the stable and unstable manifolds of x_- and x_+ have the same dimensions. □

Remark 5.6. This theorem is also applicable to *homoclinic* orbits by letting $x_- = x_+$.

Remark 5.7. If in the two-dimensional case g is Hamiltonian, $\Delta(t_0)$ coincides with the Melnikov function up to a scalar multiple, Marsden [16].

6. Heteroclinic orbits and controllability.

In this section, we show that existence of a heteroclinic solution of the unperturbed uncontrolled equation implies a controllability condition for perturbed systems with small control influence. Conversely, if the controllability condition holds for small control influence, existence of a heteroclinic solution of the unperturbed equation follows. These results are used to relate heteroclinic cycles to the existence of control sets.

Consider the following family of control systems depending on a parameter μ:

$$\dot{x} = g(x) + \mu h(x, z(t), \mu, u(t)), u \in \mathcal{U}, \tag{6.1}$$

with continuous functions g and h and control range $U \subset \mathbb{R}^m$ containing the origin; the functions z are in the hull $\mathcal{Z}$ of a single almost periodic function. We refer to $\dot{x} = g(x)$ and $\dot{x} = g(x) + \mu h(t, x, \mu, 0)$ as the unperturbed uncontrolled system and the perturbed uncontrolled system, respectively. For fixed μ this is a special case of the control system (1.1); we use the notation introduced in sections 2, 3, and 4 with a superfix μ to indicate dependence on this parameter. In particular, solutions (whose existence we always assume) are denoted by $\varphi^\mu(t, x_0, z, u)$, $t \in \mathbb{R}$, $x_0 \in \mathbb{R}^d$, $z \in \mathcal{Z}$, and $u \in \mathcal{U}$.

Proposition 6.1. *Assume that system* (6.1) *with control* $u = 0$ *satisfies the assumptions* (i)–(v) *of Theorem* 5.5. *Let* $\zeta^\mu_\pm$ *be the almost periodic solutions near the hyperbolic equilibria* $x_\pm$ *of the unperturbed uncontrolled system, and let* $x(t,\mu) := \varphi^\mu(t, x^\mu, z_0, 0)$ *be the solution near the heteroclinic orbit* ζ *from* x_- *to* x_+ *for some* $x^\mu \in \mathbb{R}^d, z_0 \in \mathcal{Z}$. *Let* μ *be a parameter value such that the conclusions of Theorem* 5.5 *hold, and assume that there are* $\varepsilon = \varepsilon(\mu), T = T(\mu) > 0$ *such that for every* $(x,z) \in Q := \operatorname{cl} V \times \mathcal{Z}$

$$\mathbf{B}_\varepsilon(\varphi^\mu(T, x, z, 0)) \subset \mathcal{O}^{\mu,+}_T(x; z, Q) \text{ and } \mathbf{B}_\varepsilon(\varphi^\mu(-T, x, z, 0)) \subset \mathcal{O}^{\mu,-}_T(x; z, Q). \tag{6.2}$$

Then there are a control function $u^\mu \in \mathcal{U}$ *and times* $t^\mu_- < t^\mu_+$ *such that the corresponding solution* $\varphi^\mu(t, x^\mu, z_0, u^\mu)$ *of* (6.1) *satisfies*

$$\varphi^\mu(t, x^\mu, z_0, u^\mu) = \begin{cases} \zeta^\mu_-(t) & \text{if } t \le t^\mu_-, \\ \zeta^\mu_+(t) & \text{if } t \ge t^\mu_+. \end{cases}$$

Proof. Pick μ as stated, and denote the constants from condition (6.2) by $\varepsilon, T > 0$. The solution $x(t,\mu)$ for the uncontrolled system satisfies (5.4). In particular, there are times $t^\mu_- < 0 < t^\mu_+$, arbitrarily large, such that

$$\left\| x(t^\mu_-, \mu) - \zeta^\mu_-(t^\mu_-) \right\| < \varepsilon \text{ and } \left\| x(t^\mu_+, \mu) - \zeta^\mu_+(t^\mu_+) \right\| < \varepsilon.$$

Together with (6.2) and the cocycle property, this means

$$\begin{aligned} \zeta^\mu_-\left(t^\mu_-\right) &\in \mathbf{B}_\varepsilon\left(\varphi^\mu\left(t^\mu_-, x^\mu, z_0, 0\right)\right) \\ &= \mathbf{B}_\varepsilon\left(\varphi^\mu\left(-T, \varphi^\mu\left(t^\mu_- + T, x^\mu, z_0, 0\right), z_0\left(t^\mu_- + T + \cdot\right), 0\right)\right) \\ &\subset \mathcal{O}^{\mu,-}_T\left(\varphi^\mu\left(t^\mu_- + T, x^\mu, z_0, 0\right); z_0\left(t^\mu_- + T + \cdot\right), Q\right) \end{aligned}$$

and, analogously,

$$\begin{aligned} \zeta^\mu_+\left(t^\mu_+\right) &\in \mathbf{B}_\varepsilon\left(\varphi^\mu\left(t^\mu_+, x^\mu, z_0, 0\right)\right) \\ &= \mathbf{B}_\varepsilon\left(\varphi^\mu\left(T, \varphi^\mu\left(t^\mu_+ - T, x^\mu, z_0, 0\right), z_0\left(t^\mu_+ - T + \cdot\right), 0\right)\right) \\ &\subset \mathcal{O}^{\mu,+}_T\left(\varphi^\mu\left(t^\mu_+ - T, x^\mu, z_0, 0\right); z_0\left(t^\mu_+ - T + \cdot\right), Q\right). \end{aligned}$$

This ensures the existence of control functions $u_{\pm}^{\mu} \in \mathcal{U}$ satisfying

$$\begin{aligned}\zeta_{-}^{\mu}\left(t_{-}^{\mu}\right) &= \varphi\left(-T, \varphi^{\mu}\left(t_{-}^{\mu}+T, x^{\mu}, z_0, 0\right), z_0\left(t_{-}^{\mu}+T+\cdot\right), u_{-}^{\mu}\right),\\ \zeta_{+}^{\mu}\left(t_{+}^{\mu}\right) &= \varphi\left(T, \varphi^{\mu}\left(t_{+}^{\mu}-T, x^{\mu}, z_0, 0\right), z_0\left(t_{+}^{\mu}-T+\cdot\right), u_{+}^{\mu}\right).\end{aligned}$$

Setting

$$u_{\mu}(t) := \begin{cases} u_{-}\left(t-t_{-}^{\mu}-T\right) & \text{if } t \in \left[t_{-}^{\mu}, t_{-}^{\mu}+T\right],\\ u_{+}\left(t-t_{+}^{\mu}+T\right) & \text{if } t \in \left[t_{+}^{\mu}-T, t_{+}^{\mu}\right],\\ 0 & \text{otherwise}\end{cases}$$

completes the proof. □

The previous proposition shows that existence of a heteroclinic orbit for the unperturbed uncontrolled equation implies the existence of a control steering the system with almost periodic excitation from the almost periodic solution near one equilibrium to the almost periodic solution near the other equilibrium. The following result considers a converse situation where the unperturbed equation has equilibria x_+ and x_- and we want to conclude from existence of controlled trajectories of the perturbed system from points near x_- to x_+ that a heteroclinic orbit of the unperturbed equation exists.

PROPOSITION 6.2. *Suppose that g and $h(x, z(t), \mu, 0)$ satisfy assumptions* (i) *and* (ii) *of Theorem* 5.5 *for all $z \in \mathcal{Z}$; i.e., these assumptions hold for system* (6.1) *with $u = 0$. Moreover, assume that the chain recurrent set of the unperturbed uncontrolled system $\dot{x} = g(x)$ relative to* $\operatorname{cl} V$ *is equal to $\{x_+, x_-\}$.*

Suppose furthermore that the control range U is bounded and there are $\mu_n \to 0$, almost periodic excitations $z_n \in \mathcal{Z}$, control functions $u_n \in \mathcal{U}$, times $t_-^n < t_+^n$, and points $x_n \in \operatorname{cl} V$ such that the solution $\varphi_n(t) := \varphi^{\mu_n}(t, x_n, z_n, u_n), t \in \mathbb{R}$, of (6.1) *is contained in* $\operatorname{cl} V$ *and satisfies $\varphi_n(t_-^n) \to x_-$ and there is $\delta > 0$ with $\|\varphi_n(t) - x_-\| \geq \delta$ for all $t \geq t_+^n$ and all n.*

Then the unperturbed uncontrolled system has a heteroclinic orbit from x_- to x_+.

Proof. For every $n \in \mathbb{N}$ let $T_n \geq t_-^n$ be the largest time satisfying $\varphi_n(T_n) \in \operatorname{cl} \mathbf{B}_r(x_-)$, where $r > 0$ is chosen such that $\mathbf{B}_r(x_-) \subset \operatorname{cl} V$. We may assume the limit $\xi_0 := \lim_{n\to\infty} \varphi_n(T_n) \in \operatorname{cl} \mathbf{B}_r(x_-)$ exists. It suffices to prove that ξ_0 lies on a heteroclinic orbit in $\operatorname{cl} V$ from x_- to x_+.

By compactness of $\mathcal{Z}$, we may assume that $z_n(T_n + \cdot)$ converges to some $z^0 \in \mathcal{Z}$. In order to show that the orbit through ξ_0 lies in $\operatorname{cl} V$, fix $t \in \mathbb{R}$ and $\varepsilon > 0$. By assumption,

$$\varphi_n(T_n) = \varphi^{\mu_n}(T_n, x_n, z_n, u_n) \to \xi_0,$$

and $\mu_n h(x, z, \mu_n, u)$ converges to zero, uniformly in (x, z, u) by continuity of h and boundedness of U. Then continuous dependence on the right-hand side and the initial value implies

$$\begin{aligned}&\varphi^{\mu_n}(T_n + t, x_n, z_n, u_n)\\ &\qquad = \varphi^{\mu_n}(t, \varphi^{\mu_n}(T_n, x_n, z_n, u_n), z_n(T_n + \cdot), u_n(T_n + \cdot)) \to \varphi^0\left(t, \xi_0, z^0, 0\right).\end{aligned}$$

Hence, the orbit through ξ_0 is contained in $\operatorname{cl} V$. Since the α- and ω-limit sets of x_0 are connected and in the chain recurrent set, they consist either of x_- or x_+. Since $\varphi^{\mu_n}(T_n + t, x_n, z_n, u_n) \in \operatorname{cl} \mathbf{B}_r(x_-)$ for $t \leq 0$, it follows that the α-limit set of ξ_0 is given by x_-. Similarly, $\varphi^{\mu_n}(T_n+t,x_n,z_n,u_n)\notin \operatorname{cl}\mathbf{B}_r(x_-)$ for $t>0$ by maximality of T_n. Thus, the ω-limit set is given by x_+. □

Next, we discuss consequences of these results for control sets of systems with almost periodic excitations. Roughly, the results above imply that the existence of a heteroclinic cycle of the unperturbed uncontrolled system is equivalent to the existence of a control set containing all almost periodic solutions near the equilibria for the systems with almost periodic excitation and small control ranges.

Recall that a heteroclinic cycle of the unperturbed equation is given by a finite set $x_0,x_1,\dots,x_n=x_0$ of equilibria together with heteroclinic solutions ζ_i from x_i to x_{i+1} for $i=0,\dots,n-1$. Existence of heteroclinic cycles can be expected in the presence of symmetries.

Theorem 6.3. *Let $x_0,x_1,\dots,x_n=x_0$ be pairwise different hyperbolic equilibria of the unperturbed uncontrolled system $\dot{x}=g(x)$, and consider control system* (6.1) *with a bounded control range U containing the origin. For $|\mu|\neq 0$, small, and $z\in\mathcal{Z}$ denote the almost periodic solutions near x_i for excitation z by $\zeta_i^\mu(z)$. Assume that system* (6.1) *with $u=0$ satisfies assumptions* (i) *and* (ii) *of Theorem* 5.5 *for all $z\in\mathcal{Z}$ on an open set V containing all equilibria x_i.*

(i) *Assume that for all i there are open subsets $V_i\subset\mathbb{R}^d$ containing the equilibria $x_-=x_i$ and $x_+=x_{i+1}$ such that assumptions* (iii)–(v) *of Theorem* 5.5 *are satisfied for* (6.1) *with $u=0$, and let $x_i(t,\mu,z)=\varphi^\mu(t,x_i^\mu,z,0)$ be the solution near the heteroclinic orbit $\zeta_i(z)$ from x_i to x_{i+1} . Assume that for all sufficiently small $|\mu|\neq 0$ there are $\varepsilon_i,T_i>0$ such that for every $(x,z)\in Q_i:=\operatorname{cl}V_i\times\mathcal{Z}$*

$$\mathbf{B}_{\varepsilon_i}(\varphi^\mu(T_i,x,z,0))\subset\mathcal{O}^{\mu,+}_{T_i}(x;z,Q_i)\ \textit{and}\ \mathbf{B}_{\varepsilon_i}(\varphi^\mu(-T_i,x,z,0))\subset\mathcal{O}^{\mu,-}_{T_i}(x;z,Q_i). \tag{6.3}$$

Then for all $|\mu|\neq 0$, small, there exists a control set D^μ such that for all $z\in\mathcal{Z}$ and all i the almost periodic solutions satisfy $\zeta_i^\mu(t)\in D^\mu_{z(t+\cdot)}$ and the heteroclinic solutions satisfy $x_i(t,\mu,z)\in D^{\mu,z(t+\cdot)}$.

(ii) *Conversely, suppose for all i there are open subsets V_i containing x_i and x_{i+1} such that the chain recurrent set of the unperturbed uncontrolled system $\dot{x}=g(x)$ relative to $\operatorname{cl}V_i$ is equal to $\{x_i,x_{i+1}\}$. Furthermore, suppose that for a sequence $0\neq\mu_n\to 0$ there are control sets D^{μ_n} containing the almost periodic solutions $\zeta_i^{\mu_n}$ near x_i for almost periodic excitations $z_n\in\mathcal{Z}$. Then the unperturbed system has a heteroclinic cycle through the x_i.*

Proof.

(i) For all i, Theorem 4.7 implies that there are control sets D_i^μ such that the almost periodic solutions $\zeta_i^\mu(z)$ are contained in the interior of $D^\mu_{i,z}$. It remains to show that all D_i^μ coincide. Fix $z\in\mathcal{Z}$, and consider the almost periodic solutions $\zeta_i(z)$ near x_i (we suppress dependence on μ in our notation). By Proposition 6.1, there are $y_1\in\mathbb{R}^d$, a control function $u_1\in\mathcal{U}$, and times $t_1<t_2$ such that the corresponding solution $\varphi(t,y_1,z,u_1)$ of (6.1) satisfies

$$\varphi(t,y_1,z,u_1)=\begin{cases}\zeta_1(t) & \text{if } t\le t_1,\\ \zeta_2(t) & \text{if } t\ge t_2.\end{cases}$$

There are $y_2\in\mathbb{R}^d$, a control function $u_2\in\mathcal{U}$, and times $\tau_2>t_2$ and $t_3>\tau_2$ such that the corresponding solution $\varphi(\cdot,y_2,z,u_1)$ of (6.1) satisfies

$$\varphi(t,y_2,z,u_2)=\begin{cases}\zeta_2(t) & \text{if } t\le \tau_2,\\ \zeta_3(t) & \text{if } t\ge t_3.\end{cases}$$

Proceeding in this way and using $x_n = x_0$, one finds times $T_2 > T_1 > 0$, a point $x \in D_1^z$, and a control $u \in \mathcal{U}$ such that

$$\varphi(T_1, x, z, u) \in D_{2,z(T_1+\cdot)} \text{ and } \varphi(T_1 + T_2, x, z, u) \in D_{1,z(T_1+T_2+\cdot)},$$
$$\text{and } \psi(t, x_1, z, u) \in Q \text{ for all } t \in [0, T_1 + T_2].$$

Then Proposition 4.5 shows $D_1 = D_2$, and, repeating this argument, one concludes that all control sets D_i coincide.

(ii) The assumptions allow us to apply Proposition 6.2. Hence, for all i, the unperturbed uncontrolled system has a heteroclinic orbit from x_i to x_{i+1}. □

7. An oscillator with *M*-potential. In this section, we will apply our results to a second order system with M-potential, which models ship roll motion.

Consider the system

$$\ddot{x} + \mu\beta_1\dot{x} + \mu\beta_3\dot{x}^3 + x - \alpha x^3 = \mu z(t) + \mu u(t) \tag{7.1}$$

with positive parameters α, β_1, and β_3, a small perturbation parameter $\mu \in \mathbb{R}$, almost periodic excitations $z : \mathbb{R} \to \mathbb{R}$, and control functions $u : \mathbb{R} \to [-\rho, \rho]$ for a control radius $\rho > 0$. This model, proposed in Kreuzer and Sichermann [14], has been studied in Colonius et al. [5] without time-dependent excitation z. Note that in this application the terms $u(\cdot)$ are interpreted as time-dependent perturbations (not as controls) where only the range $[-\rho, \rho]$ is known. Here the control sets give information on the global stability behavior: An invariant control set around the origin indicates stability. If (for large perturbation amplitudes) it has merged with a variant control set and itself becomes variant, stability is lost. Hence, it is of interest to compute all control sets.

System (7.1) is a special case of system (4.6). Hence, Proposition 4.9 shows that assumption (4.9) in Theorem 4.11 is satisfied for all $\rho_2 > \rho_1 \geq 0$. Thus, every compact chain control set E^{ρ_1} is contained in the interior of a control set D^{ρ_2}, and, hence, for all up to countably many $\rho > 0$, Remark 4.12 shows that the compact chain control sets coincide with the closures of control sets.

Writing (7.1) as a first order system yields the two-dimensional perturbed Hamiltonian system

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= -x_1 + \alpha x_1^3 + \mu\left(-\beta_1 x_2 - \beta_3 x_2^3 + z(t) + u(t)\right). \end{aligned} \tag{7.2}$$

Denote by $\varphi^\mu(t, x, z, u)$ the solution of this system, and let

$$\psi^\mu(t, x, z, u) := (\varphi^\mu(t, x, z, u), \theta_t z).$$

In the unperturbed and uncontrolled case $\mu = 0$ system (7.2) has a fixed point in the origin and two hyperbolic fixed points at $(\pm 1/\sqrt{\alpha}, 0)$. The hyperbolic fixed points are connected by two heteroclinic orbits given by $x_\pm^h(t) := \pm(x_1(t), x_2(t))$, where

$$x_1(t) := \frac{1}{\sqrt{\alpha}} \tanh \frac{t}{\sqrt{2}},\ x_2(t) := \frac{1}{\sqrt{2\alpha}} \operatorname{sech}^2 \frac{t}{\sqrt{2}},\ t \in \mathbb{R};$$

cf. Simiu [22, p. 131]. In the perturbed uncontrolled case $u \equiv 0$ denote by $\Delta_\pm$ the Melnikov functions of system (7.2) with respect to $x_\pm^h$ and denote by $\zeta_\pm^\mu$ the almost periodic solutions near $(\pm 1/\sqrt{\alpha}, 0)$, which exist for sufficiently small μ (see Proposition 5.4). Let $z_0 \in \mathcal{Z}$ be the corresponding excitation and $\xi^\mu_\pm(t) := (\zeta^\mu_\pm(t), \theta_t z_0)$.

PROPOSITION 7.1. *Assume that the almost periodic excitation z is continuously differentiable with bounded derivative. If the functions $\Delta_\pm$ have simple zeros and μ is small enough, then system* (7.2) *has a control set D containing $\xi^\mu_\pm(\mathbb{R})$. Then D will be called a heteroclinic control set.*

Proof. The proof essentially follows from Proposition 6.1. To be precise, system (7.2) satisfies assumptions (i)–(v) of Theorem 5.5 for $u = 0$: Assumption (i) is satisfied for every bounded open set $V \subset \mathbb{R}^d$ and every $\bar{\mu} > 0$. Property (ii) is clearly satisfied, because z does not depend on x and μ. Assumptions (iii) and (iv) are true for a suitable bounded and open set $V \subset \mathbb{R}^d$. Property (v) holds by assumption.

Furthermore, property (6.2) is satisfied, as can be shown by Proposition 4.9. So for sufficiently small μ Proposition 6.1 implies the existence of points $x^\mu_\pm \in \mathbb{R}^2$, control functions $u^\mu_\pm \in \mathcal{U}$, and times $s^\mu_\pm < t^\mu_\pm$ such that

$$\varphi^\mu(t, x^\mu_-, z_0, u^\mu_-) = \begin{cases} \zeta^\mu_+(t) & \text{if } t \le s^\mu_-, \\ \zeta^\mu_-(t) & \text{if } t \ge t^\mu_- \end{cases}$$

and

$$\varphi^\mu(t, x^\mu_+, z_0, u^\mu_+) = \begin{cases} \zeta^\mu_-(t) & \text{if } t \le s^\mu_+, \\ \zeta^\mu_+(t) & \text{if } t \ge t^\mu_+. \end{cases}$$

The set $\tilde{D} := \psi^\mu(\mathbb{R}, x^\mu_-, z_0, u^\mu_-) \cup \psi^\mu(\mathbb{R}, x^\mu_+, z_0, u^\mu_+) \cup \xi^\mu_-(\mathbb{R}) \cup \xi^\mu_+(\mathbb{R})$ satisfies properties (i) and (ii) of control sets and is, thus, contained in a control set D. This implies $\xi^\mu_\pm(\mathbb{R}) \subset \tilde{D} \subset D$. □

First, we study the periodic case and choose $z(t) := F \cos \omega t$ for positive parameters F and ω, which leads to the system

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= -x_1 + \alpha x_1^3 + \mu \left(-\beta_1 x_2 - \beta_3 x_2^3 + F \cos \omega t + u(t) \right). \end{aligned} \tag{7.3}$$

The excitation z is C^1 and its derivative is bounded, so Proposition 7.1 is applicable. The Melnikov functions $\Delta_\pm$ of system (7.3) can easily be computed using the residue theorem:

$$\Delta_\pm(t_0) = -\frac{2\sqrt{2}\beta_1}{3\alpha} - \frac{8\sqrt{2}\beta_3}{35\alpha^2} \pm \frac{\sqrt{2}\pi\omega F}{\sqrt{\alpha}\sinh\frac{\pi\omega}{\sqrt{2}}} \cdot \cos \omega t_0.$$

The Melnikov functions $\Delta_\pm$ have simple zeros if and only if F exceeds a certain *critical amplitude* F_c, i.e., if $F > F_c := A^{-1}B$ for

$$A := \frac{\sqrt{2}\pi\omega}{\sqrt{\alpha}\sinh\frac{\pi\omega}{\sqrt{2}}} \text{ and } B := \frac{2\sqrt{2}\beta_1}{3\alpha} + \frac{8\sqrt{2}\beta_3}{35\alpha^2}.$$

COROLLARY 7.2. *If $F > F_c$, system* (7.3) *has a heteroclinic control set for sufficiently small μ.*

Proof. The proof follows from Proposition 7.1. □

As the excitation is T-periodic for $T := 2\pi/\omega$, it is possible to compute fibers of control sets by looking at the discrete control system given by the time-T map. For the

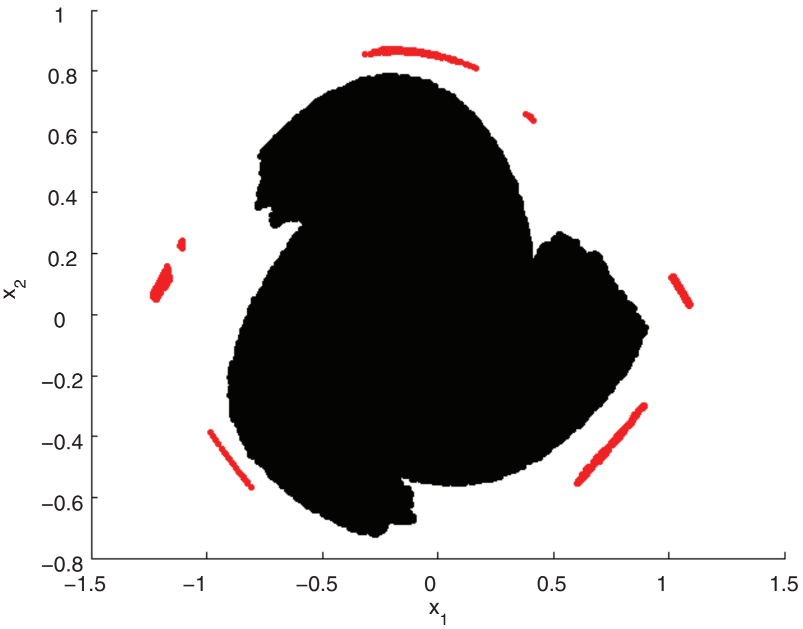


FIG. 7.1. *Fiber of control sets for the periodically excited system* (7.3). *Computed in phase* 0 *for* $\alpha = 0.674$, $\beta_1 = 0.231$, $\beta_3 = 0.375$, $\omega = 2.5$, $\rho = 1.0$, $F = 6$, *and* $\mu = 0.1$.

following computations we restrict our view to the parameter values $\alpha = 0.674$, $\beta_1 = 0.231$, and $\beta_3 = 0.375$ (see [14] for a discussion of these parameters and this choice) and choose $\omega = 2.5$ and $\rho = 1.0$. Then $F_c \approx 5.62880$, so let $F := 6 > F_c$. Figure 7.1 shows the fiber in phase 0 for $\mu = 0.1$. The control sets were approximated with the *graph algorithm* (see Dellnitz and Junge [7] and Szolnoki [24]) using the implementation in Global Analysis of Invariant Objects (GAIO).[1] For a spatial discretization into boxes, this algorithm computes strongly connected components of an associated graph whose nodes are given by the boxes and whose edges indicate reachability. The union of the resulting boxes is an approximation to a chain control set; as noted above, for system (7.1) the chain control sets typically coincide with the closures of control sets. Note that this figure shows the fiber of two control sets: an invariant control set around the origin (black) and the heteroclinic control set (red). Compare this to Figure 7.2, where the stable and unstable manifolds for these parameter values are shown, again for $\mu = 0.1$ and in phase 0.

Next, we examine quasi-periodic excitations of the form $z(t) := F\cos\omega_1 t + F\sin\omega_2 t$ for positive parameters F, ω_1, and ω_2, which leads to the system

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= -x_1 + \alpha x_1^3 + \mu\left(-\beta_1 x_2 - \beta_3 x_2^3 + F\cos\omega_1 t + F\sin\omega_2 t + u(t)\right). \end{aligned} \tag{7.4}$$

The excitation z again is C^1, and its derivative is bounded. The Melnikov functions $\Delta_\pm$ of system (7.4) are

$$\Delta_\pm(t_0) = -\frac{2\sqrt{2}\beta_1}{3\alpha} - \frac{8\sqrt{2}\beta_3}{35\alpha^2} \pm \frac{\sqrt{2}\pi F}{\sqrt{\alpha}}\left(\frac{\omega_1\cos\omega_1 t_0}{\sinh\frac{\pi\omega_1}{\sqrt{2}}} + \frac{\omega_2\sin\omega_2 t_0}{\sinh\frac{\pi\omega_2}{\sqrt{2}}}\right).$$

[1] http://www-math.uni-paderborn.de/˜agdellnitz/gaio/

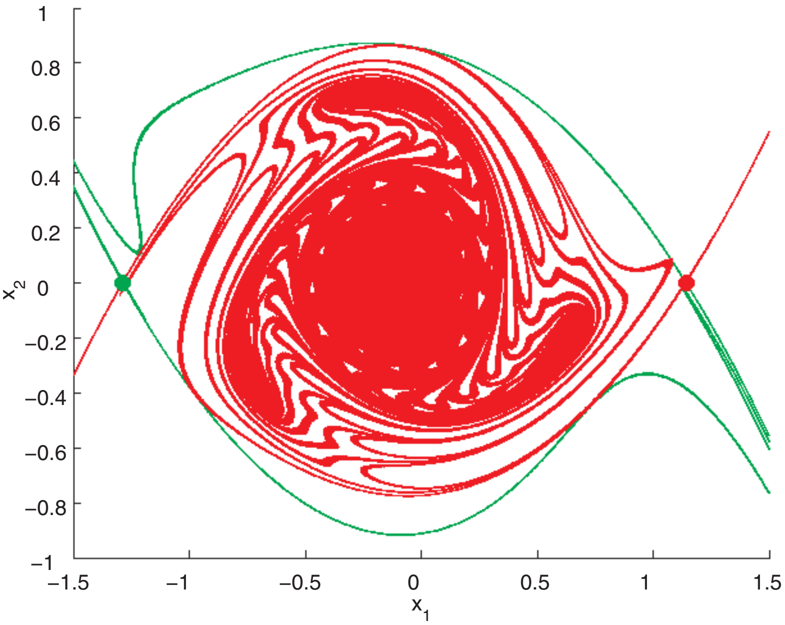


FIG. 7.2. *Stable and unstable manifolds for the uncontrolled periodically excited system* (7.3). *Computed in phase* 0 *for* $\alpha = 0.674$, $\beta_1 = 0.231$, $\beta_3 = 0.375$, $\omega = 2.5$, $F = 6$, *and* $\mu = 0.1$.

The Melnikov function $\Delta_{\pm}$ has a simple zero if $F > F_c := A^{-1}(S_1 + S_2)^{-1}B$ for

$$A := \frac{\sqrt{2}\pi}{\sqrt{\alpha}}, S_i := \frac{\omega_i}{\sinh \frac{\pi\omega_i}{\sqrt{2}}}, i = 1, 2, \text{ and } B := \frac{2\sqrt{2}\beta_1}{3\alpha} + \frac{8\sqrt{2}\beta_3}{35\alpha^2}.$$

COROLLARY 7.3. *If* $F > F_c$, *system* (7.4) *has a heteroclinic control set for sufficiently small* μ.

Proof. The proof follows from Proposition 7.1. □

Remark 7.4. The main interest in this result comes from the relations between the deterministic system and a related stochastic system, where $u(t)$ is replaced by a stochastic perturbation. Then the invariant control sets correspond to the supports of invariant measures (see, e.g., Colonius, Gayer and Kliemann [3]). For small perturbation amplitudes, system (7.1) has an invariant control set around the origin and, hence, small random perturbations will not lead to capsizing (i.e., there are no unbounded solutions $x(t)$ starting near the origin). For large perturbation amplitudes, there is no invariant control set and capsizing will occur with probability 1. Hence, it is of interest to analyze how invariance is lost. The results above indicate that this happens when the invariant control set around the origin unites with the heteroclinic control set. This shows that the picture is more complicated than indicated in [10] (where, as a simplified model, the escape equation with a single hyperbolic equilibrium was discussed).

REFERENCES

[1] H. AMANN, *Ordinary Differential Equations. An Introduction to Nonlinear Analysis*, de Gruyter Stud. Math. 13, de Gruyter, Berlin, 1990.

[2] H. BOHR, *Almost Periodic Functions*, Chelsea, New York, 1947.

[3] F. Colonius, T. Gayer, and W. Kliemann, *Near invariance for Markov diffusion systems*, SIAM J. Appl. Dyn. Syst., 7 (2008), pp. 79–107.

[4] F. Colonius and W. Kliemann, *The Dynamics of Control*, Birkhäuser, Boston, 2000.

[5] F. Colonius, E. Kreuzer, A. Marquardt, and W. Sichermann, *A numerical study of capsizing: Comparing control set analysis and Melnikov's method*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 18 (2008), pp. 1503–1514.

[6] W. Coppel, *Dichotomies in Stability Theory*, Lecture Notes in Math. 269, Springer-Verlag, Berlin, 1978.

[7] M. Dellnitz and O. Junge, *Set oriented numerical methods for dynamical systems*, in Handbook of Dynamical Systems Vol. 2, B. Fiedler, ed., Elsevier, Amsterdam, 2002, pp. 221–264.

[8] J. Falzarano, S. Shaw, and A. Troesch, *Application of global methods for analyzing dynamical systems to ship rolling motion and capsizing*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 2 (1992), pp. 101–116.

[9] A. Fink, *Almost Periodic Differential Equations*, Lecture Notes in Math. 377, Springer-Verlag, Berlin, 1974.

[10] T. Gayer, *Controllability and invariance properties of time-periodic systems*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 15 (2005), pp. 1361–1375.

[11] J. K. Hale, *Ordinary Differential Equations*, 2nd ed., Krieger, Huntington, NY, 1980.

[12] S.-R. Hsieh, A. Troesch, and S. Shaw, *A nonlinear probabilistic method for predicting vessel capsizing in random beam seas*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 446 (1994), pp. 195–211.

[13] R. A. Johnson and M. Nerurkar, *Stabilization and random linear regulator problem for linear nonautonomous control processes*, J. Math. Anal. Appl., 197 (1996), pp. 608–629.

[14] E. J. Kreuzer and W. M. Sichermann, *Investigation of large amplitude roll motions and capsizing*, in Proceedings of the Ninth International Symposium on Practical Design of Ships and Other Floating Structures, Lübeck-Travemünde, Germany, H. Keil and E. Lehmann, eds., Seehafen Verlag, 2004, pp. 689–696.

[15] B. Levitan and V. Zhikov, *Almost Periodic Functions and Differential Equations*, Cambridge University Press, Cambridge, 1982.

[16] J. E. Marsden, *Chaos in dynamical systems by the Poincaré-Melnikov-Arnold method*, in Chaos in Nonlinear Dynamical Systems, J. Chandra, ed., SIAM, Philadelphia, 1984, pp. 19–31.

[17] K. R. Meyer and G. R. Sell, *Melnikov transforms, Bernoulli bundles and almost periodic perturbations*, Trans. Amer. Math. Soc., 314 (1989), pp. 63–105.

[18] K. J. Palmer, *Exponential dichotomies and transversal homoclinic points*, J. Differential Equations, 55 (1984), pp. 225–256.

[19] L. A. B. San Martin and M. Patrao, *Morse decompositions of semiflows on fiber bundles*, Discrete Contin. Dyn. Syst., 17 (2007), pp. 113–139.

[20] N. Scherbina, *Continuity of one-parameter families of sets*, Dokl. Akad. Nauk SSSR, 234 (1977), pp. 327–329 (in Russian).

[21] J. Scheurle, *Chaotic solutions of systems with almost periodic forcing*, Z. Angew. Math. Phys., 37 (1986), pp. 12–26.

[22] E. Simiu, *Chaotic Transitions in Deterministic and Stochastic Dynamical Systems*, Princeton University Press, Princeton, NJ, 2002.

[23] M. Soliman and J. Thompson, *Basin organization prior to a tangled saddle-node bifurcation*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 1 (1991), pp. 107–118.

[24] D. Szolnoki, *Set oriented methods for computing reachable sets and control sets*, Discrete Contin. Dyn. Syst. Ser. B, 3 (2003), pp. 361–382.

[25] J. Taubert, *Kontrollmengen um Homokline Trajektorien*, Diplomarbeit, Institut für Mathematik, Universität Augsburg, Augsburg, Germany, 2006.

[26] J. M. T. Thompson, *Chaotic phenomena triggering the escape from a potential well*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 421 (1989), pp. 195–225.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 2, pp. 1080–1088

© 2009 Society for Industrial and Applied Mathematics

SEMIGROUPS OF AFFINE GROUPS, CONTROLLABILITY OF AFFINE SYSTEMS AND AFFINE BILINEAR SYSTEMS IN $\mathbf{Sl(2,\mathbb{R}) \rtimes \mathbb{R}^2}$*

OSVALDO G. DO ROCIO†, ALEXANDRE J. SANTANA†, AND MARCOS A. VERDI†

Abstract. Let $G = B \rtimes V$ be an affine group given by a semidirect product of connected Lie group B and vector space V, where B has a transitive representation on V. We study semigroups of G with nonempty interior. As an application, we obtain a characterization of controllability of bilinear systems in the case of $\mathrm{Sl}(2,\mathbb{R}) \rtimes \mathbb{R}^2$.

Key words. controllability, affine systems, semigroups, semisimple Lie group

AMS subject classifications. 93B05, 22E15, 20M20

DOI. 10.1137/080716736

1. Introduction.

In the early 1980s, Bonnard, Jurdjevic, Kupka, and Sallet (see [1], [6], and [7]) studied affine controllability (transitivity) in connection with linear controllability (transitivity). They considered an affine control system of the form

$$\dot{x} = Ax + a + \sum_{i=1}^{m} u_i(t)(B_i x + b_i), x \in \mathbb{R}^n,$$

where $A, B_1, \ldots, B_m$ are $n \times n$ matrices, $a, b_1, \ldots, b_m$ are vectors in $\mathbb{R}^n$, and $u(t) = (u_1(t), \ldots, u_m(t))$ is a control function with values in some control constraint set $\Omega \subset \mathbb{R}^n$. Hence, the study of such systems was made via analysis of a family of vector fields in the semidirect product of a Lie subgroup (generated by the exponential of $A, B_1, \ldots, B_m$) and $\mathbb{R}^n$. Now we make a short summary about one of the main results of [1], [6], and [7]. First, we recall some concepts. If $\mathcal{F}$ is a family of complete vector fields on a manifold M and S is a semigroup generated by $\cup_{X\in\mathcal{F}}\{\exp tX : t \geq 0\}$, then $\mathcal{F}$ is said to be transitive or controllable if $Sx = M$ for each $x \in M$, where Sx is the orbit of S through x. A vector field $X(x) = Ax + a$, where A is an $n \times n$ matrix and a is a vector in $\mathbb{R}^n$, is called an affine field. An arbitrary family $\mathcal{F}$ of affine fields in $\mathbb{R}^n$ is called an affine family. If $X(x) = Ax + a$ is an affine field, $\overrightarrow{X}$ denotes the corresponding linear field $x \mapsto Ax$ for all $x \in V$. Any affine family $\mathcal{F}$ defines the corresponding linear family $\overrightarrow{\mathcal{F}} = \{\overrightarrow{X}; X \in \mathcal{F}\}$. In this context, it was proved that if the affine family has no fixed point (that is, for each $x \in V$, there exists $X \in \mathcal{F}$ such that $X(x) \neq 0$), then the affine controllability depends only on its linear controllability. More precisely, in [7], the authors proved the following.

THEOREM 1.1. *If $\mathcal{F}$ is an affine family on vector space V such that the corresponding linear family $\overrightarrow{\mathcal{F}}$ is transitive on $V \setminus \{0\}$ and if $\mathcal{F}$ has no fixed points in V, then $\mathcal{F}$ is transitive on V.*

In order to write the above theorem in the context of semigroups of Lie groups, it is necessary to introduce some notations and definitions. Let V be an n-dimensional

*Received by the editors February 26, 2008; accepted for publication (in revised form) November 14, 2008; published electronically March 6, 2009.
http://www.siam.org/journals/sicon/48-2/71673.html

†Departamento de Matemática, Universidade Estadual de Maringá, Cx. Postal 331, 87.020-900, Maringá, Pr, Brasil (rocio@uem.br, ajsantana@uem.br, maverdi@uem.br). The first two authors were supported by Fundação Araucária grant 6240.

real vector space. Denote by End$V$ the set of all linear endomorphism on V and by $\mathrm{Gl}(V)$ the set of all automorphisms on V. Consider a subgroup H of $\mathrm{Gl}(V)$, with transitive action on $V \setminus \{0\}$ (i.e., $Hx = V \setminus \{0\}$ for all $x \in V \setminus \{0\}$) and take the group $G = H \rtimes V$ given by the semidirect product of H and V. Recall that the affine group operation is defined by $(g, v) \cdot (h, w) = (gh, v + gw)$ for all $(g, v), (h, w) \in G$. Let $\pi : G \to H$ be the canonical projection of the affine group on Lie group H. The action of G on V, given by $(g, v) \cdot w = gw + v$, with $(g, v) \in G$ and $w \in V$, is called affine action. The natural action of $\pi(G) = H$ on V is called linear action. Given a semigroup $S \subset G$, the affine action of S on V is said transitive on V (or simply transitive) if $Sx = V$ for all $x \in V$ and the linear action of semigroup $T \subset H$ is called transitive on $V \setminus \{0\}$ (or simply transitive) if $Tx = V \setminus \{0\}$ for all $x \in V \setminus \{0\}$. Moreover, $v \in V$ is called fixed point under S if $Sv = \{v\}$. We can now restate the previous theorem.

THEOREM 1.2. *Consider affine group $G = H \rtimes V$. Let $S \subset G$ be a connected semigroup with nonempty interior. Suppose that the linear action of $\pi(S)$ is transitive on $V \setminus \{0\}$ and that S has no fixed point. Then the affine action of S on V is transitive.*

In the introduction of [7], the authors mentioned an example of an affine system $\mathcal{F}$, which has no fixed point in $\mathbb{R}^n$ and whose linear part $\vec{\mathcal{F}}$ is transitive on $\mathbb{R}^n \setminus \{0\}$ but for which the set of accessibility of the corresponding right invariant system is a proper semigroup of $\mathrm{Sl}(n, \mathbb{R}) \rtimes \mathbb{R}^n$. However, a careful analysis in this example shows that the corresponding semigroup is not a subsemigroup of $\mathrm{Sl}(n, \mathbb{R}) \rtimes \mathbb{R}^n$. One of the purposes of the present paper is to show that it is not possible to obtain such an example in $\mathrm{Sl}(n, \mathbb{R}) \rtimes \mathbb{R}^n$. In this context, our main result is as follows.

THEOREM 1.3. *Let $G = B \rtimes V$ be an affine group, where B is a semisimple Lie group that acts transitively on $V \setminus \{0\}$. Let $S \subset G$ be a connected semigroup with nonempty interior. Suppose that the linear action of $\pi(S)$ is transitive on $V \setminus \{0\}$. Then the affine action of S on V is transitive.*

Note that, as a consequence of this theorem, the semigroup S has no fixed point. Then it follows that the algebraic condition of the vector fields to generate the Lie algebra (rank condition) substitutes the geometric condition of fixed points. Due to the relation between the rank condition and the condition int$S \neq \emptyset$, it is natural to consider the last one in the context of our work.

This paper is organized as follows. In the next two sections, we prove that if semigroup $\pi(S)$ is linear transitive on V, then the hypothesis that S has no fixed point is not necessary. Contrary to [1], [6], and [7], we use another way, that is, the main tools used in our proof are contained in the theory of semigroups in a semisimple Lie group (in particular, the following Theorem 3.1) and in the general theory of semigroups (in particular, the purity lemma). In the last section, as an application of the above result, we study affine control systems of the type

$$\dot{x} = Ax + a + uBx + ub, \tag{1.1}$$

where $A, B \in \mathfrak{sl}(2, \mathbb{R})$, $a, b \in \mathbb{R}^2$, and $u \in \mathbb{R}$. Briefly, consider affine group $G = \mathrm{Sl}(2, \mathbb{R}) \rtimes \mathbb{R}^2$ and take the subsemigroup S of G generated by $\exp(t(A, a))$ and $\exp(\gamma(B, b))$, where $t, \gamma \in \mathbb{R}$, and $t \geq 0$. Denote $A_1 = -2\mathrm{tr}(AB)I + 2AB - BA$, $A_2 = 4\det(B)I + B^2$, $B_1 = 4\det(A)I + A^2$, and $B_2 = 2\mathrm{tr}(AB)I - 2BA + AB$. We prove that if $\det([A, B]) \neq 0$, then the Lie algebra generated by (A, a) and (B, b) coincides with the Lie algebra of $G = \mathrm{Sl}(2, \mathbb{R}) \rtimes \mathbb{R}^2$ if and only if $A_1a - B_1b \neq 0$ or $A_2a + B_2b \neq 0$. With this result and Theorem 5.3 in Braga Barros et al. [4], we conclude the following result: given a transitive system as in (1.1), with $\det([A,B]) < 0$ and $A_1a - B_1b \neq 0$ or $A_2a + B_2b \neq 0$, then the above system is controllable.

2. The case where $\pi(S)$ is a transitive group. Let H be a connected Lie group. In this section, we consider $G = H \rtimes V$ an affine group given by the semidirect product of H and finite-dimensional vector space V, where H has a transitive representation on V.

Take $\pi : G \to H$ the canonical projection. Note that, given a semigroup $S \subset G$ with nonempty interior, we have that $\pi(S)$ is a subsemigroup of H with nonempty interior. In this section, we consider the case where $\pi(S)$ is a group, that is, $\pi(S) = H$. Hence, we have the main result of this section.

Theorem 2.1. *Let $G = H \rtimes V$ be an affine group. Denote by $S \subset G$ a connected semigroup with nonempty interior. Supposing that $\pi(S) = H$ is transitive on $V \setminus \{0\}$, we have $S = G$. In particular, the affine action of S on V is transitive.*

Some of the following assertions are known (see, e.g., [1] and [6]), but we make a brief mention of them in the context of semigroups.

Consider the normal abelian subgroup of G

$$N = \{(1,v) \text{ such that } v \in V\}.$$

As N is normal, it is not difficult to show that SN is a semigroup of G.

Now with the above hypothesis, we suppose, on the contrary, that $S \neq G$, that is, S is a proper semigroup with nonempty interior. In this case, S is contained in a maximal semigroup with nonempty interior in G, and hence, there is no loss of generality in assuming that S is a maximal proper semigroup of G. We note that the set

$$T = \{(1,v) \text{ such that } (1,v) \in \mathrm{int}S\}$$

is a nonempty abelian semigroup. In fact, as $\mathrm{int}S \neq \emptyset$, there exists $(g,v) \in \mathrm{int}S$. Then $g \in \pi(S)$, and as we are assuming that $\pi(S)$ is a group, it follows that $g^{-1} \in \pi(S)$. Hence, there exists $w \in V$ such that $(g^{-1}, w) \in S$. As $\mathrm{int}S$ is an ideal, we have $(g,v)(g^{-1},w) = (1, gw + v) \in \mathrm{int}S$, and therefore, T is nonempty.

One important fact here is that for each $(1,v) \in T$ and $g \in B$, there exists an integer $n \in \mathbb{N}$ such that $(1, ngv) \in T$. In particular, there exists an integer $n \in \mathbb{N}$ such that $(1, n(-v)) \in T$. In fact, for each $g \in H = \pi(S)$, there exist $v_1, v_2 \in V$ such that $(g, v_1), (g^{-1}, v_2) \in S$. Then $(g^{-1}, v_2)(g, v_1) = (1, g^{-1}v_1 + v_2) \in S$. As $(1,v) \in \mathrm{int}S$, there exists $n \in \mathbb{N}$ such that $(1, v - n^{-1}(g^{-1}v_1 + v_2)) \in \mathrm{int}S$. Then $(1, v - n^{-1}(g^{-1}v_1 + v_2))^n = (1, nv - (g^{-1}v_1 + v_2)) \in \mathrm{int}S$. Hence,

$$(g, v_1)\left(1, nv - \left(g^{-1}v_1 + v_2\right)\right)\left(g^{-1}, v_2\right) = (1, ngv) \in \mathrm{int}S.$$

Therefore, $(1, ngv) \in T$. To verify the particular case, note that using the transitivity of $\pi(S)$ on $V \setminus \{0\}$, it follows that there exists $(g, w) \in S$ such that $gv = -v$. Hence, $(1, n(-v)) \in T$.

Now we need the following lemma to conclude the proof of Theorem 2.1.

Lemma 2.2. *If $(1,w) \in S$, then $(1, \mathbb{R}^+ w) \subset S$.*

Proof. Take $(1,w) \in S$. Knowing that

$$(1,w) = \left(1, \frac{1}{n}w\right)^n \in N = \{(1,v) \text{ such that } v \in V\},$$

it follows that $(1, \frac{1}{n}w)^n = (1, w) \in S$ for all integers $n > 1$. As we are supposing that S is a maximal semigroup and as N is a normal abelian subgroup of G, we have, by the purity lemma (see, e.g., Lemma v.5.10 in Hilgert–Hofmann–Lawson [5]), that $(1, \frac{1}{n}w) \in S$. Then $(1, \frac{1}{n}w)^m = (1, \frac{m}{n}w) \in S$ for $\frac{m}{n} > 0$. Hence, as S is connected, it follows that $(1, \mathbb{R}^+ w) \subset S$.

Having disposed of the previous results, we can now return to $(1, u) \in \operatorname{int}S$. If $(1, -u) \in S$, then $(1, 0) \in \operatorname{int}S$, which contradicts the fact that S is a proper semigroup. Then $(1, u) \in \operatorname{int}S$, but $(1, -u) \notin S$. Note that, by the above comments, there exists an integer $n \in \mathbb{N}$ such that $(1, n(-u)) \in \operatorname{int}S$, and, by Lemma 2.2, it follows that $(1, \mathbb{R}^+ n(-u)) \subset S$. Then

$$\left(1, \frac{1}{n} n(-u)\right) = (1, -u) \in S.$$

As $(1, u) \in$ int$S$, then $(1, u)(1, -u) = (1, 0) \in \operatorname{int}S$. Therefore, $S = G$, contradicting our assumption that $S \neq G$. This completes the proof of Theorem 2.1. □

3. The case where *H* is a semisimple Lie group. In this section, we consider the case where $H = B$ is a connected, semisimple, and noncompact Lie group with finite center. Suppose that B has a transitive representation on V. We recall that the classification of these transitive Lie groups can be found in Boothby [2] and Kramer [8].

The next theorem is fundamental in this paper, and it is a summary of some concepts and results of the theory of semigroups in a semisimple Lie group. Let B be a connected, semisimple, and noncompact Lie group with finite center. Take V a finite-dimensional vector space V, where B has a transitive representation on V. Take $S \subset B$ a semigroup with nonempty interior. With these hypotheses, we have what follows.

THEOREM 3.1. *The semigroup S of B is transitive on $V \setminus \{0\}$ if and only if $S = B$.*

Proof. Proposition 5.6 of San Martin [9] shows that S is controllable in the n-dimensional space V if and only if S is controllable on the projective space $\mathbb{P}^{n-1}$, here controllability is equivalent to transitivity (see Proposition 4.4 in [9]). The cited proposition refers to subgroups of the linear groups which are transitive on finite-dimensional vector space V and hence, on the projective space. Then, for every transitive group B, we have diffeomorphism $B/P \simeq \mathbb{P}^l$, to a correspondent l given by the transitive representation of B on V, where $P \subset B$ is the isotropy subgroup of the canonical element of $\mathbb{P}^l$. Then the conditions of Theorem 1.2 of San Martin and Tonelli [10] are satisfied. Therefore, if S is transitive on $B/P \simeq \mathbb{P}^l$ or equivalently, on the l-dimensional vector space V, we have $S = B$. □

Consider again $G = B \rtimes V$ an affine group given by a semidirect product of B and a finite-dimensional vector space V and $\pi : G \to B$ the canonical projection. In this context, we have the following theorem.

THEOREM 3.2. *Let $G = B \rtimes V$ be an affine group. Denote by $S \subset G$ a connected semigroup with nonempty interior. Supposing that the linear action of $\pi(S)$ is transitive on $V \setminus \{0\}$, we have that the affine action of S on V is transitive.*

Proof. Since $\pi(S)$ is a nonempty semigroup of group B and as the linear action of $\pi(S)$ on $V \setminus \{0\}$ is transitive, we have by Theorem 3.1 that $\pi(S) = B$ is a transitive group. Therefore, by Theorem 2.1, the affine action of S on V is transitive. □

4. Controllability of affine bilinear systems in $\mathbf{Sl(2,\mathbb{R}) \rtimes \mathbb{R}^2}$. The purpose of this section is to present a detailed analysis of the controllability of affine bilinear systems

$$\dot{x} = Ax + a + uBx + ub, \tag{4.1}$$

with unrestricted control $u \in \mathbb{R}$, where $A, B \in \mathfrak{sl}(2, \mathbb{R})$ and $a, b \in \mathbb{R}^2$. More precisely, consider affine group $G = \mathrm{Sl}(2, \mathbb{R}) \rtimes \mathbb{R}^2$ and its Lie algebra $\mathfrak{sl}(2, \mathbb{R}) \rtimes \mathbb{R}^2$. Note that for simplicity, we use the same product symbol for the group and algebra. Let S be the subsemigroup of G generated by $\exp(t(A, a))$ and $\exp(\gamma(B, b))$, where $t, \gamma \in \mathbb{R}$ and $t \geq 0$. We find conditions on (A, a) and (B, b) such that $\mathrm{int}S \neq \emptyset$ and $Sx = \mathbb{R}^2$ for all $x \in \mathbb{R}^2$, i.e., conditions for the controllability of the above system.

To find these conditions, we use the following fact proved in [4]: $\pi(S)$ is transitive on $\mathbb{R}^2 \setminus \{0\}$ if and only if $\det[A, B] < 0$. In particular, in this situation, we have that $\{A, B, [A, B]\}$ is a basis for $\mathfrak{sl}(2, \mathbb{R})$. Moreover, it is well known that S has nonempty interior if and only if the Lie subalgebra of $\mathfrak{sl}(2, \mathbb{R}) \rtimes \mathbb{R}^2$, generated by (A, a), (B, b), coincides with $\mathfrak{sl}(2, \mathbb{R}) \rtimes \mathbb{R}^2$. Recall that the Lie subalgebra generated by (A, a), (B, b) is the smallest subalgebra of $\mathfrak{sl}(2, \mathbb{R}) \rtimes \mathbb{R}^2$ containing the set $\{(A, a)(B, b)\}$. With these facts and using an algorithm of transitivity (see Boothby and Wilson [3]), we give conditions on (A, a) and (B, b) for the controllability of the above system.

Let $\mathfrak{g}$ be a five-dimensional Lie algebra and take two linearly independent elements $X, Y \in \mathfrak{g}$. By this algorithm of transitivity, in order to know if a Lie subalgebra generated by $X, Y \in \mathfrak{g}$ is $\mathfrak{g}$, it is sufficient to analyze just the following sets:

$$\begin{aligned}
\beta_1 &= \{X, Y, [X, Y], [X, [X, Y]], [Y, [X, Y]]\},\\
\beta_2 &= \{X, Y, [X, Y], [X, [X, Y]], [X, [X, [X, Y]]]\},\\
\beta_3 &= \{X, Y, [X, Y], [X, [X, Y]], [Y, [X, [X, Y]]]\},\\
\beta_4 &= \{X, Y, [X, Y], [X, [X, Y]], [[X, Y], [X, [X, Y]]]\},\\
\beta_5 &= \{X, Y, [X, Y], [Y, [X, Y]], [X, [Y, [X, Y]]]\},\\
\beta_6 &= \{X, Y, [X, Y], [Y, [X, Y]], [Y, [Y, [X, Y]]]\},\\
\beta_7 &= \{X, Y, [X, Y], [Y, [X, Y]], [[X, Y], [Y, [X, Y]]]\}.
\end{aligned}$$

In fact, let $\mathfrak{g}_1$ be the vector subspace of $\mathfrak{g}$ spanned just by X and Y and take its basis $\{X, Y\}$. We argue by induction. For $k > 1$, define $\mathfrak{g}_k$ as the subspace of $\mathfrak{g}$ spanned by $\mathfrak{g}_{k-1}$ together with the collection of all the elements of the form $[Z, W]$, where $Z, W \in \mathfrak{g}_{k-1}$. It follows that $\mathfrak{g}_k$ is spanned by a set containing a basis β of $\mathfrak{g}_{k-1}$ and the brackets of elements of β. Hence, we can select a basis of $\mathfrak{g}_k$ formed by X, Y, and its successive brackets. The process finishes when $\mathfrak{g}_k = \mathfrak{g}_{k-1}$ or $\dim \mathfrak{g}_k = 5$. Note that, in this case, are necessary at most four steps to find a basis to the Lie subalgebra spanned by $\{X, Y\}$. Moreover, this basis is one of those listed above.

Now we take $X = (A, a), Y = (B, b) \in \mathfrak{sl}(2, \mathbb{R}) \rtimes \mathbb{R}^2$. Recall that the bracket $[(A, a), (B, b)]$ is given by

$$[(A, a), (B, b)] = ([A, B], Ab - Ba).$$

Now we show that is enough to consider just the sets β_2, β_3, β_5, and β_6 in order to verify if the Lie algebra generated by X and Y coincides with $\mathfrak{sl}(2, \mathbb{R}) \rtimes \mathbb{R}^2$. More precisely, we prove that if at least one of the vectors $A_1 a - B_1 b$ or $A_2 a - B_2 b$, described below, is not null, then at least one of these sets is a basis of $\mathfrak{sl}(2, \mathbb{R}) \rtimes \mathbb{R}^2$.

We begin by describing the successive brackets in terms of A and B.

LEMMA 4.1. *Suppose that* $\det[A, B] \neq 0$. *Then:*

1. $[A, [A, B]] = -2\mathrm{tr}(AB)A - 4\det(A)B$.

2. $[B,[A,B]] = 4\det(B)A + 2\mathrm{tr}(AB)B$.

The proof of this lemma can be verified by a direct computation. Hence, using the bilinearity property and Jacobi's identity, we can express the other brackets in terms of A,B and $[A,B]$. For instance, $[A,[A,[A,B]]] = -4\det(A)[A,B]$ and $[B,[A,[A,B]]] = 2\mathrm{tr}(AB)[A,B]$. For simplicity, we will use the following notation:

$$\begin{aligned} A_1 &= -2\mathrm{tr}(AB)I + 2AB - BA,\\ A_2 &= 4\det(B)I + B^2,\\ B_1 &= 4\det(A)I + A^2,\\ B_2 &= 2\mathrm{tr}(AB)I - 2BA + AB, \end{aligned}$$

where I stands for the identity matrix 2×2.

Now we establish necessary conditions for β_2, β_3, β_5, or β_6 to be a basis for $\mathfrak{sl}(2,\mathbb{R}) \rtimes \mathbb{R}^2$.

LEMMA 4.2. *Assume that* $\det[A,B] \neq 0$. *Then*

1. $\{(A,a),(B,b),([A,B],Ab-Ba),([A,[A,B]],A(Ab-Ba)-[A,B]a)\}$ *is linearly independent if and only if*

$$A_1 a - B_1 b \neq 0.$$

2. $\{(A,a),(B,b),([A,B],Ab-Ba),([B,[A,B]],B(Ab-Ba)-[A,B]b)\}$ *is linearly independent if and only if*

$$A_2 a + B_2 b \neq 0.$$

Proof. By Corollary 5.2 in [4], we have that $\{A,B,[A,B]\}$ is a linearly independent set. Hence,

$$\{(A,a),(B,b),([A,B],(Ab-Ba))\}$$

is also a linearly independent set. Thus,

$$\{(A,a),(B,b),([A,B],Ab-Ba),([A,[A,B]],A(Ab-Ba)-[A,B]a)\}$$

is linearly dependent if and only if the last element is a linear combination of the others, that is, if there exist $\alpha,\beta,\gamma \in \mathbb{R}$ such that

$$\alpha A + \beta B + \gamma[A,B] = [A,[A,B]]$$

and

$$\alpha a + \beta b + \gamma(Ab - Ba) = A(Ab-Ba) - [A,B]a.$$

By Lemma 4.1, the first of these equalities is equivalent to $\alpha = -2\mathrm{tr}(AB)$, $\beta = -4\det(A)$, and $\gamma = 0$. Therefore,

$$\{(A,a),(B,b),([A,B],Ab-Ba),([A,[A,B]],A(Ab-Ba)-[A,B]a)\}$$

is linearly dependent if and only if

$$-2\mathrm{tr}(AB)a - 4\det(A)b = A(Ab-Ba) - [A,B]a = A^2b - 2ABa + BAa.$$

This concludes the proof of item (1). The proof of (2) is analogous. □

In the next lemma, we assume $A_1a - B_1b \neq 0$ and show that β_2 and β_3 are not basis for $\mathfrak{sl}(2,\mathbb{R}) \rtimes \mathbb{R}^2$ if and only if $A_1a - B_1b$ is an eigenvector of A and B. This result will be important in the proof of Proposition 4.5.

LEMMA 4.3. *Suppose that* $\det([A,B]) \neq 0$ *and* $A_1a - B_1b \neq 0$. *Then*

1. $\{(A,a),(B,b),([A,B],Ab-Ba),([A,[A,B]],A(Ab-Ba)-[A,B]a),$
$([A,[A,[A,B]]],A(A(Ab-Ba)-[A,B]a)-[A,[A,B]]a)\}$
is linearly independent if and only if

$$A(A_1a-B_1b)\neq\delta(A_1a-B_1b) \textit{ for all } \delta\in\mathbb{R}.$$

2. $\{(A,a),(B,b),([A,B],Ab-Ba),([A,[A,B]],A(Ab-Ba)-[A,B]a),$
$([B,[A,[A,B]]],B(A(Ab-Ba)-[A,B]a)-[A,[A,B]]b)\}$
is linearly independent if and only if

$$B(A_1a-B_1b)\neq\delta(A_1a-B_1b) \textit{ for all } \delta\in\mathbb{R}.$$

Proof. By hypothesis,

$$\{(A,a),(B,b),([A,B],Ab-Ba),([A,[A,B]],A(Ab-Ba)-[A,B]a)\}$$

is linearly independent. Thus, the set in item (1) is linearly dependent if and only if there exists $\alpha,\beta,\gamma,\delta\in\mathbb{R}$ such that

$$\alpha A+\beta B+\gamma[A,B]+\delta[A,[A,B]]=[A,[A,[A,B]]]$$

and

$$\alpha a+\beta b+\gamma(Ab-Ba)+\delta(A(Ab-Ba)-[A,B]a)$$
$$=A(A(Ab-Ba)-[A,B]a)-[A,[A,B]]a.$$

By Lemma 4.1, the first of these equalities is equivalent to $\alpha=2\delta\mathrm{tr}(AB)$, $\beta=4\delta\det(A)$, and $\gamma=-4\det(A)$. Replacing these values in the second equation and using, again, Lemma 4.1, we have

$$2\delta\mathrm{tr}(AB)a+4\delta\det(A)b-4\det(A)(Ab-Ba)+\delta(A^2b-2ABa+BAa)$$
$$=A(A^2b-2ABa+BAa)+(2\mathrm{tr}(AB)A+4\det(A)B)a.$$

Hence, the proof of item (1) is concluded by simplifying the expressions. The second item is proved in an analogous way. □

Similarly, we have the following result.

Lemma 4.4. *Suppose that* $\det([A,B])\neq 0$ *and* $A_2a+B_2b\neq 0$. *Then*

1. $\{(A,a),(B,b),([A,B],Ab-Ba),([B,[A,B]],B(Ab-Ba)-[A,B]b),$
$([A,[B,[A,B]]],A(B(Ab-Ba)-[A,B]b)-[B,[A,B]]a)\}$
is linearly independent if and only if

$$A(A_2a+B_2b)\neq\delta(A_2a+B_2) \textit{ for all } \delta\in\mathbb{R}.$$

2. $\{(A,a),(B,b),([A,B],Ab-Ba),([B,[A,B]],B(Ab-Ba)-[A,B]b),$
$([B,[B,[A,B]]],B(B(Ab-Ba)-[A,B]B)-[B,[A,B]]b)\}$
is linearly independent if and only if

$$B(A_2a+B_2b)\neq\delta(A_2a+B_2b) \textit{ for all } \delta\in\mathbb{R}.$$

In the next proposition, we will need of the following general fact about Lie algebras: Let X and Y be elements of a finite-dimensional Lie algebra $\mathfrak{g}$ and suppose that $\{X,Y,[X,Y]\}$ is a linearly independent set. If $\{X,Y,[X,Y],[X,[X,Y]]\}$ and

$\{X, Y, [X, Y], [Y, [X, Y]]\}$ are both linearly dependent, then the Lie algebra generated by X and Y has dimension three. In fact, let X_1 and X_2 be elements of vector space V generated by X, Y, and $[X, Y]$ and write $X_1 = a_1X + b_1Y + c_1[X, Y]$, $X_2 = a_2X + b_2Y + c_2[X, Y]$. Then

$$\begin{aligned}[X_1, X_2] &= [a_1X + b_1Y + c_1[X, Y], a_2X + b_2Y + c_2[X, Y]] \\ &= a_1b_2[X, Y] + a_1c_2[X, [X, Y]] - b_1a_2[X, Y] + b_1c_2[Y, [X, Y]] \\ &\quad - c_1a_2[X, [X, Y]] - c_1b_2[Y, [X, Y]].\end{aligned}$$

Now the hypothesis implies that $[X, [X, Y]]$ and $[Y, [X, Y]]$ are both linear combination of X, Y and $[X, Y]$. Hence, $[X_1, X_2] \in V$, concluding the proof.

With these results, we obtain the following proposition.

PROPOSITION 4.5. *Suppose that* $\det([A, B]) \neq 0$. *Then the Lie algebra generated by* (A, a) *and* (B, b) *coincides with* $\mathfrak{sl}(2, \mathbb{R}) \rtimes \mathbb{R}^2$ *if and only if* $A_1a - B_1b \neq 0$ *or* $A_2a + B_2b \neq 0$.

Proof. Suppose that $A_1a - B_1b \neq 0$. Note that if at least one of the items of Lemma 4.3 is satisfied, then the Lie algebra generated by (A, a) and (B, b) coincides with $\mathfrak{sl}(2, \mathbb{R}) \rtimes \mathbb{R}^2$. If we suppose the contrary, then A and B have a common eigenvector. Thus, there is a basis for $\mathbb{R}^2$ such that A and B are triangular superior matrices with respect to this basis. Then $\det([A, B]) = 0$, which contradicts the hypothesis. Analogously, if $A_2a + B_2b \neq \{0\}$, then at least one of the items of Lemma 4.4 is satisfied. Therefore, if $A_1a - B_1b \neq 0$ or $A_2a + B_2b \neq 0$, then the Lie algebra generated by (A, a) and (B, b) is $\mathfrak{sl}(2, \mathbb{R}) \rtimes \mathbb{R}$. To the converse, consider $A_1a - B_1b = 0$ and $A_2a + B_2b = 0$, then, by Lemma 4.2, sets $\{(A, a), (B, b), ([A, B], Ab - Ba), ([A, [A, B]], A(Ab - Ba) - [A, B]a)\}$ and $\{(A, a), (B, b), ([A, B], Ab - Ba), ([B, [A, B]], B(Ab - Ba) - [A, B]b)\}$ are both linearly dependent. Thus, by the commentary that precedes this proposition, we conclude that the Lie algebra generated by (A, a) and (B, b) has dimension three. □

With this proposition, Theorem 5.3 in [4] and Theorem 3.2 we get the following.

THEOREM 4.6. *Consider the above assumptions. If* $\det([A, B]) < 0$ *and* $A_1a - B_1b \neq 0$ *or* $A_2a + B_2b \neq 0$, *then system* 4.1 *is controllable.*

Now consider system 4.1 with a special matrix B. Using the above theorem, we give conditions (on A, B, a, and b) for the controllability of the system.

Example. If $\det(B) > 0$, then $B = \begin{pmatrix} 0 & d \\ -d & 0 \end{pmatrix}$ with respect to some basis of $\mathbb{R}^2$. Suppose that $A = \begin{pmatrix} r & s \\ t & -r \end{pmatrix}$ in this basis. Thus, if $a = (x, y)$ and $b = (z, w)$, then

$$\det([A, B)] = -d^2\left(2st + s^2 + t^2 + 4r^2\right),$$

$$A_1a - B_1b = 3\left(-tdx + rdy + \left(r^2 + st\right)z, rdx - sdy + \left(r^2 + st\right)w\right)$$

and

$$A_2a + B_2b = 3\left(d^2x - sdz + rdw, d^2y + rdz + tdw\right).$$

The third equality implies that $A_2a + B_2b = 0$ if and only if

$$x = \frac{sz - rw}{d} \text{ and } y = -\frac{rz + tw}{d}.$$

But, replacing these values in the second equality, we obtain $A_1a - B_1b = 0$. Therefore, using the above theorem, it follows that if

$$\det([A,B]) = -d^2\left(2st + s^2 + t^2 + 4r^2\right) < 0 \text{ and}$$

$$A_2a + B_2b \neq 0 \left(\Leftrightarrow x \neq \frac{sz - rw}{d} \text{ or } y \neq -\frac{rz + tw}{d}\right),$$

then system 4.1 is controllable. As a specific instance, take $A = \left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$ and $B = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$. Since $\det([A,B]) = -4$, we have that if $x \neq -w$ or $y \neq -z$, then system 4.1 is controllable.

In the case of $\det(B) < 0$ and $\det(B) = 0$, we obtain similar conditions.

Acknowledgments. We are grateful to anonymous referees whose careful and competent review helped to improve the paper.

REFERENCES

[1] B. Bonnard, V. Jurdjevic, I. Kupka, and G. Sallet, *Transitivity of families of invariant vector fields on the semidirect products of Lie groups*, Trans. Amer. Math. Soc., 271 (1982), pp. 525–535.

[2] W. M. Boothby, *A transitivity problem from control theory*, J. Differential Equations, 17 (1975), pp. 296–307.

[3] W. M. Boothby and E. N. Wilson, *Determination of the transitivity of bilinear systems*, Siam J. Control Optim., 17 (1979), pp. 212–221.

[4] C. J. Braga Barros, J. G. Ribeiro, O. G. Rocio, and L. A. B San Martin, *Controllability of two dimensional bilinear systems*, Proyecciones, 15 (1996), pp. 111–139.

[5] J. Hilgert, K. Hofmann, and J. Lawson, *Lie Groups, Convex Cones and Semigroups*, Oxford University Press, London, 1989.

[6] V. Jurdjevic and G. Sallet, *Controllability of affine systems*, Differential Geometric Control Theory, R. Brockett, R. Millman and H. Sussman, eds., Birkauser, Boston, 1982, pp. 299–309.

[7] V. Jurdjevic and G. Sallet, *Controllability properties of affine systems*, SIAM J. Control Optim., 22 (1984), pp. 501–508.

[8] L. Kramer, *Two-transitive Lie groups*, Journal Reine Angew. Math., 563 (2003), pp. 83–113.

[9] L. A. B. San Martin, *Invariant control sets on flag manifolds*, Math. Control Signals Systems, 6 (1993), pp. 41–61.

[10] L. A. B. San Martin and P. A. Tonelli, *Transitive actions of semigroups in semi-simple Lie groups*, Semigroup Forum, 58 (1999), pp. 142–151.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 2, pp. 1089–1111

© 2009 Society for Industrial and Applied Mathematics

A NONSTANDARD APPROACH TO A DATA ASSIMILATION PROBLEM AND TYCHONOV REGULARIZATION REVISITED[*]

JEAN-PIERRE PUEL[†]

Abstract. We consider evolution problems, such as diffusion convection equations or the linearized Navier–Stokes system, or a weak coupling of them, which we would like to "predict" on a time interval $(T_0, T_0 + T)$ but for which, on one hand, the initial value of the state variable is unknown. On the other hand "measurements" of the solutions are known on a time interval $(0, T_0)$ and, for example, on a subdomain in the space variable. The classical approach in variational data assimilation is to look for the initial value at time 0, and this is known to be an ill-posed problem which has to be regularized. Here we propose to look for the value of the state variable at time T_0 (the end time of the "measurements") and we prove on some basic examples that this is a well-posed problem. We give a result of exact reconstruction of the value at T_0 which is based on global Carleman inequalities, and we give an approximation algorithm which uses classical optimal control auxiliary problems. Using the same mathematical arguments, we also show why Tychonov regularization for variational data assimilation works in practical situations corresponding to realistic applications.

Key words. data assimilation, Carleman estimates, evolution equations

AMS subject classifications. 93C41, 93B05, 49N30, 49N45

DOI. 10.1137/060670961

1. Introduction.

Data assimilation problems are of great importance for practical purposes, in particular for meteorological prediction, ocean models, and environmental sciences. They lead to very heavy computations, and to give an idea of this importance (for example, in meteorological prediction), we mention that the computation time dedicated to data assimilation corresponds to more than one half of the total computation time. An example of these problems can be roughly, but simply, described as follows (cf. also an interesting description in [1]). The phenomenon under consideration (for example, meteorological prediction) is modeled by a (very complex, in general) system of evolution equations which can be written in the form

$$\frac{\partial Y}{\partial t} + \mathcal{A}Y + \mathcal{N}Y = F, \tag{1.1}$$

where Y is the vector representing the state variables that we want to "predict," $\mathcal{A}$ is a partial differential (elliptic) operator in space variables which can be assumed to be linear, $\mathcal{N}$ is a nonlinear operator which is in general of lower order, and F is the (known) vector of exterior forces which act on the system. Our goal is to compute a (good) approximation of Y during a period of time of length T (prediction). Of course many people have been working on these questions in practice and have developed various efficient (even rather complex) methods of discretization for system (1.1), but they are confronted by the following problem: they do not know the "initial data" for Y at a given time before T_0 in order to compute the solution of the prediction model from T_0 on. However, they know "measurements" of Y in some spatial regions

[*]Received by the editors September 28, 2006; accepted for publication (in revised form) October 18, 2008; published electronically March 11, 2009. A preliminary version of these results has appeared in [21].
http://www.siam.org/journals/sicon/48-2/67096.html

[†]Laboratoire de Mathématiques de Versailles, Université de Versailles Saint-Quentin, 45 Avenue des Etats Unis, 78035 Versailles, France (jppuel@math.uvsq.fr).

during a time interval $(0, T_0)$. The data assimilation problem is then to determine an approximation of initial data at a time before T_0 from the known "measurements." A great number of both theoretical and practical works, as well as computations, have been devoted to these questions. For articles related to the problem under consideration here we can refer to [15], [16], [2], and the references therein.

In the classical approach the problem is considered as follows. In order to fix ideas we shall say that T_0 is today, 0 is yesterday, and $T_0 + T$ is tomorrow. We know "measurements" of Y on a time interval $(0, T_0)$ (from yesterday to today). We look for the initial value $Y(0)$ (value of the state variable yesterday) in order to "compute" Y on the time interval $(0, T_0 + T)$ (from yesterday to tomorrow). This problem is known to be ill-posed, and one of the methods generally used to solve it is what is called variational data assimilation (see [1], [15], [16], [2]). This method uses optimal control techniques for minimizing a suitable cost function (usually a least square method) on a linearized problem together with a regularization method (for example, Tychonov regularization) and an optimality system making use of the adjoint state.

First, we present a new approach which is based on controllability techniques and which is nonclassical in the following sense. We know the "measurements" of Y on $(0, T_0)$. What is really important for prediction purposes is to be able to "compute" Y on the time interval $(T_0, T_0 + T)$ (from today until tomorrow). We will then look for an approximation of $Y(T_0)$ (and no approximations of $Y(0)$). Our interest in this approach is that we will prove, on significant examples of linear problems (unfortunately, not on a realistic model of meteorology which would be, by far, too complex) that this problem is well-posed, and we can give a good estimate of its sensitivity to errors in the measurements. We will also give an exact method of reconstruction of $Y(T_0)$ and an approximate method which turns out to be simpler in practice, in particular for numerical computations, and for which we will prove convergence towards the exact solution.

The underlying mathematical techniques are those used for exact controllability to trajectories and are based on global Carleman inequalities.

In the next section we give two basic examples of equations, diffusion convection equations and linearized Navier–Stokes equations, which can be rigorously treated. The case of weakly coupled systems, such as linearized Boussinesq equations, can also be rigorously proved, but we do not present this case here in order to clarify the presentation. It is then straightforward to derive a general principle which can be applied to a large class of evolution systems. This can be done formally or nonformally, depending on the possibility of proving a global Carleman estimate for the system under consideration.

Second, we study Tychonov regularization in a way which is different from the classical works on the subject (see, for example, [10] or [4]), and we try to explain why it works in practical situations for variational data assimilation methods without abstract assumptions on the data. In fact we show that, due to Carleman estimates, the minimization problem (without regularization), which is ill-posed in the variational data assimilation context (where the control variable is the initial condition $Y(0)$), is well-posed in a class of trajectories of the problem which may not have an initial value. Then, assuming a hypothesis on the regularity of the minimizer, which can be viewed as natural in realistic models (the minimizer is here a trajectory and not an initial value), we show that the solution of the problem with Tychonov regularization converges to this minimizer. Using an additional hypothesis which is similar to (yet different than) the one used, for example, in [10], [5], or [4], we give a convergence rate for the previous convergence.

2. Nonstandard approach on two basic examples.

2.1. Diffusion convection equations. Let Ω be a bounded open subset of $\mathbb{R}^N$ of class C^2 with boundary Γ, and consider the following diffusion convection equation:

$$(2.1)\quad \frac{\partial y}{\partial t} - \sum_{i,j=1}^{N} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial y}{\partial x_j}\right) + \sum_{i=1}^{N} b_i \frac{\partial y}{\partial x_i} + \sum_{j=1}^{N} \frac{\partial}{\partial x_j}(c_j y) = f \text{ in } \Omega \times (0, T_0),$$

$$(2.2)\quad y = 0 \text{ on } \Gamma \times (0, T_0).$$

Notice that we do not impose any initial condition on y and that we do not know the initial value $y(0)$. We assume that, for example,

(2.3)

$$f \in L^2(0, T_0; L^2(\Omega)), \quad a_{ij} \in W^{1,\infty}(\Omega \times (0, T_0)),\ a_{ij} = a_{ji}, \quad b_i, c_j \in L^\infty(\Omega \times (0, T_0)),$$

and the coefficients a_{ij} satisfy an ellipticity condition

$$(2.4)\qquad \exists \gamma > 0\ \ \forall (x,t) \in \Omega \times [0, T_0],\ \forall \xi \in \mathbb{R}^N,\ \sum_{i,j=1}^{N} a_{ij}(x,t)\xi_i\xi_j \geq \gamma|\xi|^2.$$

Let ω be a nonempty open subset of Ω, and let χ_ω denote the characteristic function of ω. We suppose that we know a "measurement" of the solution

$$(2.5)\qquad y.\chi_\omega = h$$

with

$$(2.6)\qquad h \in L^2(0, T_0; L^2(\omega)).$$

Remark 2.1. It will be very important to make precise the class of functions in which y satisfying (2.1) and (2.2) is taken. This will be a key point in what follows.

Remark 2.2. Here, for simplicity, we suppose that we know a "measurement" of the solution y in the interior of Ω. Without many changes, one could consider the case where "measurements" of the flux $\frac{\partial y}{\partial \nu_A}$ are known on a nonempty part Γ_0 of the boundary Γ, where $\nu = (\nu_1, \ldots, \nu_N)$ is the outward unit normal on Γ, and

$$\frac{\partial y}{\partial \nu_A} = \sum_{i,j=1}^{N} a_{ij}\frac{\partial y}{\partial x_j}\nu_i.$$

Let us first briefly recall what is usually done in variational data assimilation (see, for example, [1], [15], [16]).

For $y_0 \in L^2(\Omega)$ (which will be the control variable) we consider the solution of (2.1), (2.2) satisfying in addition

$$(2.7)\qquad y(0) = y_0 \text{ in } \Omega.$$

It is then well known that problem (2.1), (2.2), (2.7) has a unique solution $y = y[y_0]$ with

$$y \in C([0, T_0]; L^2(\Omega)) \cap L^2(0, T_0; H_0^1(\Omega)).$$

In what follows, we will always consider this regularity when speaking of solutions of problem (2.1), (2.2), (2.7). This will no longer be the case when we drop the initial condition (2.7).

Let us now consider the cost functional

$$J(y_0) = \frac{1}{2}\int_0^{T_0}\int_\omega |y[y_0] - h|^2 dxdt. \tag{2.8}$$

(We could have taken a different functional expressing the error between the observation on $y[y_0]$ and the measurement h.)

We would like to solve the following minimization problem:

$$\begin{aligned} &\text{Find } \bar{y}_0 \in L^2(\Omega) \text{ such that} \\ &J(\bar{y}_0) = \min_{y_0 \in L^2(\Omega)} J(y_0). \end{aligned} \tag{2.9}$$

It is well known that this problem is ill-posed as, if we take a minimizing sequence y_0^n for J, we cannot obtain any estimate on this sequence (in any known functional space).

Therefore it is usual to add a Tychonov regularization to our functional; namely, we consider for $\alpha > 0$ the new functional

$$J_\alpha(y_0) = \frac{1}{2}\int_0^{T_0}\int_\omega |y[y_0] - h|^2 dxdt + \frac{\alpha}{2}|y_0|^2_{L^2(\Omega)}, \tag{2.10}$$

and we now want to solve the following optimal control problem:

$$\begin{aligned} &\text{Find } y_\alpha \in L^2(\Omega) \text{ such that} \\ &J_\alpha(y_\alpha) = \min_{y_0 \in L^2(\Omega)} J_\alpha(y_0). \end{aligned} \tag{2.11}$$

It is then standard, using classical optimal control methods (see [17]), to show that this problem has a unique solution. But of course the functional has been changed, and it is not clear what the solution y_α represents. The main questions are then what happens when $\alpha \to 0$ and what is the sensitivity to perturbations in the measurements h. We will give a partial answer to these questions later on.

We now come to a nonstandard approach to the data assimilation problem using controllability techniques.

In order to state our result we need to introduce the adjoint (backward) problem, where we allow a control on the region in which the measurement is given.

For $\varphi_0 \in L^2(\Omega)$ and $v \in L^2(0, T_0; L^2(\omega))$ we denote by φ the solution of

$$-\frac{\partial\varphi}{\partial t} - \sum_{i,j=1}^N \frac{\partial}{\partial x_j}\left(a_{ji}\frac{\partial\varphi}{\partial x_i}\right) - \sum_{i=1}^N \frac{\partial}{\partial x_i}(b_i\varphi) - \sum_{j=1}^N c_j\frac{\partial\varphi}{\partial x_j} = v.\chi_\omega \text{ in } \Omega \times (0, T_0), \tag{2.12}$$

$$\varphi = 0 \text{ on } \Gamma \times (0, T_0), \tag{2.13}$$

$$\varphi(T_0) = \varphi_0. \tag{2.14}$$

We know that this problem has a unique solution, and we know that

$$\varphi \in C([0, T_0]; L^2(\Omega)) \cap L^2(0, T_0; H_0^1(\Omega)).$$

It is by now well known (cf., for example, [8]) that we have null controllability for this problem, which means that for every $\varphi_0 \in L^2(\Omega)$, there exists $v \in L^2(0, T_0; L^2(\omega))$

such that $\varphi(0) = 0$. We have to make things more precise for our purposes. This requires some technicalities, which can be skipped by readers who are only interested in the formal method.

Let us call y^0 the solution of problem (2.1), (2.2), (2.7) with $y_0 = 0$. Then if we call y the solution of (2.1), (2.2), (2.7) with general initial condition $y_0 \in L^2(\Omega)$, we have $y = y^0 + z$, where z is a solution of (2.1), (2.2), (2.7) with $f = 0$. Let us denote

$$\mathcal{V}^0 = \{z,\ z \text{ solution of (2.1), (2.2), (2.7) with } f = 0,\ y_0 \in L^2(\Omega)\} \tag{2.15}$$

and

$$\mathcal{V} = y^0 + \mathcal{V}^0. \tag{2.16}$$

Then $\mathcal{V}^0$ is a vector space, and of course $\mathcal{V}$ is an affine space. We can now derive a global Carleman estimate for elements of $\mathcal{V}$, but this requires the introduction of some weights.

Let ω_0 be a nonempty open set such that $\overline{\omega}_0 \subset \omega$ (for example, ω_0 can be a small open ball). Then we know from [8] (see also [20] for a detailed proof) that there exists $\psi \in C^2(\overline{\Omega})$ such that

$$\begin{aligned} \psi(x) > 0 &\quad \forall x \in \Omega, \\ \psi(x) = 0 &\quad \forall x \in \Gamma, \\ |\nabla\psi(x)| \neq 0 &\quad \forall x \in \overline{\Omega} - \omega_0. \end{aligned}$$

We now use the function ψ to build new weights. Let us define the following for $\lambda > 0$ and for an integer $k \geq 1$ (here we need only $k = 1$, but for further extensions it happens that we sometimes need to take $k > 1$, which does not cause any change in what follows):

$$\xi_k(x,t) = \frac{e^{\lambda(m|\psi|_{L^\infty(\Omega)} + \psi(x))}}{t^k(T_0 - t)^k}, \tag{2.17}$$

$$\eta_k(x,t) = \frac{e^{(\frac{k+1}{k})\lambda m|\psi|_{L^\infty(\Omega)}} - e^{\lambda(m|\psi|_{L^\infty(\Omega)} + \psi(x))}}{t^k(T_0 - t)^k}, \tag{2.18}$$

where $m > k$.

We can notice that η_k tends rapidly to $+\infty$ when $t \to T_0$ or $t \to 0$ but that η_k is uniformly bounded in $\Omega \times [\delta, T_0 - \delta]$ if $0 < \delta < \frac{T_0}{2}$.

We will also need in section 3 the weights $\tilde{\eta}_k$ and $\tilde{\xi}_k$ defined by

$$\tilde{\eta}_k(t) = \eta_k(t) \text{ if } t \in \left[0, \frac{T_0}{2}\right],\ \tilde{\eta}_k(t) = \eta_k\left(\frac{T_0}{2}\right) \text{ if } t \in \left[\frac{T_0}{2}, T_0\right], \tag{2.19}$$

$$\tilde{\xi}_k(t) = \xi_k(t) \text{ if } t \in \left[0, \frac{T_0}{2}\right],\ \tilde{\xi}_k(t) = \xi_k\left(\frac{T_0}{2}\right) \text{ if } t \in \left[\frac{T_0}{2}, T_0\right]. \tag{2.20}$$

Notice that for every $\delta > 0$, $\tilde{\eta}_k$ and $\tilde{\xi}_k$ are bounded on $[\delta, T_0]$.

Our final weight will depend on a second positive parameter s and will be of the form $e^{-s\eta_1(x,t)}$. We can see that, for fixed s, this function tends very rapidly to 0 when $t \to T_0$ or $t \to 0$.

Then, following [8] or [20], where the complete proofs can be found, and the method of [14] when $c = (c_j) \neq 0$, we can state a global Carleman estimate for $y \in \mathcal{V}$.

PROPOSITION 2.3. *There exist parameters $s_0 > 0$ and $\lambda_0 > 0$ and there exists a constant $C > 0$ depending only on Ω, ω_0, ψ, and T_0, on γ defined in* (2.4)*, and on the coefficients $a_{i,j}$ such that for every $s > s_0$, for every $\lambda > \lambda_0$, and for every $y \in \mathcal{V}$, we have*

$$\begin{aligned} (2.21)\qquad & s\lambda^2 \int_0^{T_0} \int_\Omega \xi_1 e^{-2s\eta_1} |\nabla y|^2 dxdt + s^3\lambda^4 \int_0^{T_0} \int_\Omega \xi_1^3 e^{-2s\eta_1} |y|^2 dxdt \\ & \leq C\Big(\int_0^{T_0} \int_\Omega e^{-2s\eta_1} |f|^2 dxdt + s^3\lambda^4 \int_0^{T_0} \int_\omega \xi_1^3 e^{-2s\eta_1} |y|^2 dxdt \Big). \end{aligned}$$

This inequality will turn out to be fundamental. From now on we fix $s > s_0$ and $\lambda > \lambda_0$. As a first consequence of this proposition we have the following result.

PROPOSITION 2.4. *The bilinear form defined by*

$$\forall z, \tilde{z} \in \mathcal{V}^0,\ (z, \tilde{z})_{V^0} = \int_0^{T_0} \int_\omega z.\tilde{z} dxdt$$

is a scalar product on $\mathcal{V}^0$.

Remark 2.5. We could have taken a weight $\xi_1^3 e^{-2s\eta_1}$ in the expression of the above scalar product, as we can notice that

$$\exists M > 0 \text{ such that } \xi_1^3 e^{-2s\eta_1} \leq M.$$

However, this would not lead to real improvements in what follows and would make the presentation more complicated.

The proof of Proposition 2.4 requires only the following unique continuation property for $z \in \mathcal{V}^0$:

$$\forall z \in \mathcal{V}^0,\ z = 0 \text{ in } \omega \times (0, T_0)\ \Rightarrow\ z = 0 \text{ in } \Omega \times (0, T_0).$$

This is an extension (in terms of regularity) of the well-known result by Mizohata [19] and is, for example, a consequence of inequality (2.21). If $z \in \mathcal{V}^0$, this corresponds to the case $f = 0$, and (2.21) says that if $z = 0$ in $\omega \times (0, T_0)$, then the left-hand side must be zero, which ensures that $z = 0$ in the whole domain $\Omega \times (0, T_0)$. This shows Proposition 2.4.

DEFINITION 2.6. *We denote by V^0 the (abstract) completion of $\mathcal{V}^0$ with respect to the norm $|.|_{V^0}$ associated with the above defined scalar product $(.,.)_{V^0}$ and denote by V the translated space*

$$(2.22)\qquad V = y^0 + V^0.$$

We then have immediately from the completion argument the following corollary.

COROLLARY 2.7. *The space V^0 is a Hilbert space for the scalar product $(.,.)_{V^0}$, and V is the associated complete metric space. For every $y \in V$, inequality* (2.21) *still holds true.*

Remark 2.8. (1) The spaces V^0 and V can appear to be very abstract but, in fact, because of inequality (2.21), we know that V is contained in the weighted Sobolev space of L^2 functions with the weight $\xi_1^3 e^{-2s\eta_1}$ with gradients in L^2 with the weight $\xi_1 e^{-2s\eta_1}$ such that (2.1) and (2.2) (which now make perfect sense) hold true. Elements of V are therefore solutions of (2.1) and (2.2) and are also called trajectories of the problem.

(2) As the weight degenerates near $t = 0$, functions of V may not have any value at $t = 0$ in any sense (we will refer to the value at $t = 0$ as the initial value, as these functions are solutions of an evolution problem) and this is crucial to observe.

(3) The weights $\xi_1 e^{-2s\eta_1}$ and $\xi_1^3 e^{-2s\eta_1}$ are uniformly bounded from below by a positive constant in $\Omega \times [\delta, T_0 - \delta]$ if $\delta > 0$. Therefore, if $y \in V$, then

$$\forall \delta > 0,\ \ y \in L^2(\delta, T; H_0^1(\Omega)).$$

If now $\theta \in C^\infty[0, T_0]$ with

$$0 \leq \theta \leq 1,\ \ \theta(t) = 0 \text{ if } t \in \left[0, \frac{\delta}{2}\right],\ \ \theta(t) = 1 \text{ if } t \in [\delta, T_0],$$

considering the equation satisfied by $z = \theta.y$, multiplying this equation by z, and using classical energy estimates for the diffusion convection operator (see also the proof of Theorem 2.9 below), we see that

$$y \in C([\delta, T_0]; L^2(\Omega)) \cap L^2(\delta, T_0; H_0^1(\Omega)),$$

and there exists a constant $C(\delta) > 0$ such that, for every $y \in V^0$,

$$|y|^2_{C([\delta,T_0];L^2(\Omega))} + ||y||^2_{L^2(\delta,T_0;H_0^1(\Omega))} \leq C(\delta)|y|^2_{V^0} = C(\delta)\int_0^{T_0}\int_\omega |y|^2 dxdt. \tag{2.23}$$

In particular, for $y \in V$, the value $y(T_0)$ makes perfect sense in $L^2(\Omega)$.

We can now state a first result giving stability and exact reconstruction of $y(T_0)$ with respect to the "measurement" $h = y_{/\omega\times(0,T_0)}$.

THEOREM 2.9. *If Ω is a bounded open subset of $\mathbb{R}^N$ of class C^2, if the coefficients $a_{i,j}, b_i$, and c_j and the functions f and h satisfy the previous hypotheses* (2.3), (2.4), *and* (2.6) *for any nonempty $\omega \subset \Omega$, for any $T_0 > 0$, and for any $\varphi_0 \in L^2(\Omega)$, there exists $v = v(\varphi_0) \in L^2(0, T_0; L^2(\omega))$ such that the solution φ of* (2.12), (2.13), (2.14) *verifies*

$$\varphi(0) = 0. \tag{2.24}$$

Taking $v(\varphi_0)$ of minimal norm among admissible controls, the mapping $\varphi_0 \to v(\varphi_0)$ is continuous and

$$\exists C > 0\ \forall \varphi_0 \in L^2(\Omega),\ \ |v(\varphi_0)|_{L^2(0,T_0;L^2(\omega))} \leq C|\varphi_0|_{L^2(\Omega)}. \tag{2.25}$$

We then have, if $y \in V$ (which means that $h = y_{/\omega\times(0,T_0)} \in L^2(0, T_0; L^2(\omega))$),

$$\forall \varphi_0 \in L^2(\Omega),\ \int_\Omega y(T_0)\varphi_0 dx = \int_0^{T_0}\int_\Omega f\varphi dxdt - \int_0^{T_0}\int_\omega hv(\varphi_0)dxdt. \tag{2.26}$$

Moreover, there exists a constant $C > 0$ depending only on Ω, ω, and T_0 and the coefficients a_{ij}, b_i, and c_j such that

$$|y(T_0)|^2_{L^2(\Omega)} \leq C\left(\int_0^{T_0}\int_\omega |h|^2 dxdt + \int_0^{T_0}\int_\Omega |f|^2 dxdt\right). \tag{2.27}$$

Remark 2.10. Inequality (2.27) is of course independent of the *unknown* value $y(0)$ which in fact may not exist. It depends only on the right-hand side f and the "measurement" $h = y_{/\omega\times(0,T_0)}$. It is a stability inequality.

Equality (2.26) enables us to calculate the component of $y(T_0)$ on φ_0 for any $\varphi_0 \in L^2(\Omega)$ knowing the "measurement" $h = y_{/\omega\times(0,T_0)}$, the right-hand side f, and the "control" $v(\varphi_0)$, which has to be computed. Taking successively for φ_0 elements of a Hilbert basis of $L^2(\Omega)$, we can therefore reconstruct exactly $y(T_0)$. Of course, when dealing with numerical approximations, we would take φ_0 in a finite dimensional basis to obtain approximations of $y(T_0)$. The method has been used and numerical computations have been performed for a large-scale ocean model in [9], giving very promising results. Of course the use of reduced basis will be of major importance in using this method for numerical experiments. Initial work in this direction has been done in [11] and it gives very positive results.

The method allows us also to measure the sensitivity of the reconstruction of $y(T_0)$ with respect to perturbations in h, as will be seen in Corollary 2.11 below, and this is a very important feature in practice.

Of course, in real applications, the measurements are not provided in an open set but on a finite number of points, but this case seems to be impossible to treat mathematically.

Proof of Theorem 2.9. We already know that Carleman estimate (2.21) holds true for every $y \in V$. Let us now take a cut-off function $\theta \in C^\infty[0, T_0]$, such that

$$\begin{aligned} &0 \le \theta(t) \le 1 \quad \forall t \in [0, T_0],\\ &\theta(t) = 0 \quad \forall t \in \left[0, \frac{T_0}{4}\right],\\ &\theta(t) = 1 \quad \forall t \in \left[\frac{3T_0}{4}, T_0\right], \end{aligned}$$

and define $\tilde{y}(x,t) = \theta(t).y(x,t)$.

Then from (2.1), (2.2) we see that $\tilde{y}$ satisfies the following problem:

$$\begin{aligned} &\frac{\partial \tilde{y}}{\partial t} - \sum_{i,j=1}^{N} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial \tilde{y}}{\partial x_j}\right) + \sum_{i=1}^{N} b_i \frac{\partial \tilde{y}}{\partial x_i} + \sum_{j=1}^{N} \frac{\partial}{\partial x_j}(c_j \tilde{y}) = \theta f + y.\theta' \text{ in } \Omega \times (0, T_0),\\ &\tilde{y} = 0 \text{ on } \Gamma \times (0, T_0),\\ &\tilde{y}(x, 0) = 0 \text{ in } \Omega. \end{aligned}$$

Using now classical energy estimates (multiplying the equation by $\tilde{y}$) we obtain, as $\theta' = 0$ on $[0, \frac{T_0}{4}] \cup [\frac{3T_0}{4}, T_0]$ (C may denote different constants),

$$|\tilde{y}(T_0)|^2_{L^2(\Omega)} \le C\left(\int_0^{T_0}\int_\Omega |f|^2 dxdt + \int_{\frac{T_0}{4}}^{3\frac{T_0}{4}}\int_\Omega |y|^2 dxdt\right).$$

But, now $\tilde{y}(T_0) = y(T_0)$, and from (2.21), due to the fact that on $[\frac{T_0}{4}, \frac{3T_0}{4}]$ the weight $\xi_1^3 e^{-2s\eta_1}$ is bounded from below, we have

$$|y(T_0)|^2_{L^2(\Omega)} \le C\left(\int_0^{T_0}\int_\Omega |f|^2 dxdt + \int_0^{T_0}\int_\omega |y|^2 dxdt\right),$$

which is exactly the stability inequality (2.27).

But when $f = 0$ this can also be viewed as an observability inequality for (2.1) with $f = 0$, (2.2), and (2.7), which corresponds to the adjoint problem of the backward control problem (in the φ variable) (2.12), (2.13), (2.14). It is now well known (see [8] or [20], for example) that this observability inequality implies existence of a control $v = v(\varphi_0)$ such that the solution φ of (2.12), (2.13), (2.14) verifies

$$\varphi(0) = 0 \text{ in } \Omega.$$

The same observability inequality gives the continuity of the mapping $\varphi_0 \in L^2(\Omega) \to v(\varphi_0) \in L^2(0, T_0; L^2(\omega))$ when $v(\varphi_0)$ is taken to be the control of minimal norm among admissible controls. This proves inequality (2.25).

It remains to show equality (2.26). If $y \in \mathcal{V}$, then multiplying (2.1) by φ and taking into account that $\varphi(0) = 0$ we immediately obtain (2.26). Now if $y \in V$, taking a Cauchy sequence (y^n) with $y^n \in \mathcal{V}$ and $y^n \to y$ in V, we see immediately from (2.27) that $(y^n(T_0))$ is a Cauchy sequence in $L^2(\Omega)$ and $y^n(T_0) \to y(T_0)$ in $L^2(\Omega)$. Therefore we can pass to the limit in (2.26), which therefore remains true for $y \in V$. This finishes the proof of Theorem 2.9. □

The previous recovery method also gives an estimate of the sensitivity to error measurements. We have the following result.

Corollary 2.11. *We write $y(T_0)$ for the recovery obtained by the previous method from an exact measurement h and write $\hat{y}(T_0)$ for the recovery obtained using a measurement $\hat{h}$. Then there exists a constant $C > 0$, independent of h and $\hat{h}$, such that*

$$|y(T_0) - \hat{y}(T_0)|_{L^2(\Omega)} \leq C|h - \hat{h}|_{L^2(0,T_0;L^2(\omega))}. \tag{2.28}$$

Proof of Corollary 2.11. First of all we recall that the problem giving $v(\varphi_0)$ does not depend on the measurement. Then from (2.26) we immediately obtain

$$\forall \varphi_0 \in L^2(\Omega), \ \int_\Omega (y(T_0) - \hat{y}(T_0))\varphi_0 dx = \int_0^{T_0} \int_\omega (h - \hat{h}) v(\varphi_0) dx dt.$$

Therefore, taking the supremum over all $\varphi_0 \in L^2(\Omega)$ with $|\varphi_0|_{L^2(\Omega)} = 1$, we obtain

$$|y(T_0) - \hat{y}(T_0)|_{L^2(\Omega)} \leq |h - \hat{h}|_{L^2(0,T_0;L^2(\omega))} \sup_{|\varphi_0|_{L^2(\Omega)}=1} |v(\varphi_0)|_{L^2(0,T_0;L^2(\omega))}.$$

Because of (2.25), we immediately obtain (2.28) and the proof is complete. □

The previous result gives an "exact" method for reconstructing $y(T_0)$ but it relies on the resolution of null controllability problems. Hereafter we give an approximation method which makes use of more classical optimal control problems (other approximations could also be developed) and we prove convergence of these approximations.

Let us consider the following optimal control problem for fixed $\varphi_0 \in L^2(\Omega)$. Let φ be a solution of (2.12), (2.13), (2.14), and for $\beta > 0$ let us define

$$K_\beta(v) = \frac{1}{2\beta} \int_\Omega |\varphi(0)|^2 dx + \frac{1}{2} \int_0^{T_0} \int_\omega |v|^2 dx dt. \tag{2.29}$$

We look for $v_\beta \in L^2(0, T_0; L^2(\omega))$ such that

$$K_\beta(v_\beta) = \min_{v \in L^2(0,T_0;L^2(\omega))} K_\beta(v). \tag{2.30}$$

We obtain the following result.

THEOREM 2.12. (1) *For every $\beta > 0$ there exists a unique solution $v_\beta \in L^2(0,T_0;L^2(\omega))$ to problem* (2.30), *and v_β is characterized by the optimality system*

$$\begin{aligned}
&-\frac{\partial\varphi_\beta}{\partial t} - \sum_{i,j=1}^N \frac{\partial}{\partial x_j}\left(a_{ji}\frac{\partial\varphi_\beta}{\partial x_i}\right) - \sum_{i=1}^N \frac{\partial}{\partial x_i}(b_i\varphi_\beta) - \sum_{j=1}^N c_j\frac{\partial\varphi_\beta}{\partial x_j} = v_\beta\chi_\omega \text{ in } \Omega\times(0,T_0),\\
&\varphi_\beta = 0 \text{ on } \Gamma\times(0,T_0),\\
&\varphi_\beta(T_0) = \varphi_0 \text{ in } \Omega,\\
&\frac{\partial p_\beta}{\partial t} - \sum_{i,j=1}^N \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial p_\beta}{\partial x_j}\right) + \sum_{i=1}^N b_i\frac{\partial p_\beta}{\partial x_i} + \sum_{j=1}^N \frac{\partial}{\partial x_j}(c_j p_\beta) = 0 \text{ in } \Omega\times(0,T_0),\\
&p_\beta = 0 \text{ on } \Gamma\times(0,T_0),\\
&p_\beta(0) = \frac{1}{\beta}\varphi_\beta(0) \text{ in } \Omega,\\
&p_\beta + v_\beta = 0, \text{ in } \omega\times(0,T_0).
\end{aligned}$$

(2) *When β tends to zero, we have*

$$v_\beta \to \bar v \text{ in } L^2(0,T_0;L^2(\omega)),\ \varphi_\beta \to \bar\varphi \text{ in } C([0,T_0];L^2(\Omega))\cap L^2(0,T_0;H_0^1(\Omega)),$$

where $\bar v$ and $\bar\varphi$ satisfy

$$-\frac{\partial\bar\varphi}{\partial t} - \sum_{i,j=1}^N \frac{\partial}{\partial x_j}\left(a_{ji}\frac{\partial\bar\varphi}{\partial x_i}\right) - \sum_{i=1}^N \frac{\partial}{\partial x_i}(b_i\bar\varphi) - \sum_{j=1}^N c_j\frac{\partial\bar\varphi}{\partial x_j} = \bar v\chi_\omega \text{ in } \Omega\times(0,T_0), \tag{2.31}$$

$$\bar\varphi = 0 \text{ on } \Gamma\times(0,T_0), \tag{2.32}$$

$$\bar\varphi(T_0) = \varphi_0 \text{ in } \Omega, \tag{2.33}$$

and

$$\bar\varphi(0) = 0 \text{ in } \Omega. \tag{2.34}$$

Moreover, $\bar v = v(\varphi_0)$ is the element of minimal norm in $L^2(0,T_0;L^2(\omega))$ such that (2.31), (2.32), (2.33), *and* (2.34) *occur.*

Finally, when $\beta \to 0$, we have

$$\int_0^{T_0}\int_\Omega f\varphi_\beta dxdt - \int_0^{T_0}\int_\omega h.v_\beta dxdt \to \int_\Omega y(T_0)\varphi_0 dx. \tag{2.35}$$

Proof. For $\beta > 0$, (2.30) is a classical optimal control problem which is known to have a unique solution v_β (see, for example, [17]) and the characterization given in (1) is standard.

We also know from Theorem 2.9 that there exists v such that (2.31), (2.32), (2.33), and (2.34) are satisfied. It is clear that the set of such elements v is a nonempty closed convex set in $L^2(0,T_0;L^2(\omega))$ so that there exists a unique $\bar v \in L^2(0,T_0;L^2(\omega))$, which minimizes the norm in this set, and we have $\bar v = v(\varphi_0)$, where $v(\varphi_0)$ is as given by Theorem 2.9. The corresponding solution of (2.31), (2.32), (2.33) will be denoted by $\bar\varphi$.

Now $\bar v$ is admissible in (2.30) and, as $\bar\varphi(0) = 0$, we have for every $\beta > 0$,

$$K_\beta(v_\beta) \le K_\beta(\bar v) = \frac{1}{2}\int_0^{T_0}\int_\omega |\bar v|^2 dxdt.$$

This shows that when $\beta \to 0$, we have

$$\int_0^{T_0}\int_\omega |v_\beta|^2 dxdt \le \int_0^{T_0}\int_\omega |\bar{v}|^2 dxdt,$$
$$\frac{1}{\beta}\int_\Omega |\varphi_\beta(0)|^2 dx \le \int_0^{T_0}\int_\omega |\bar{v}|^2 dxdt.$$

Consequently we can extract a subsequence, still denoted by (v_β), such that

$$v_\beta \rightharpoonup \tilde{v} \text{ in } L^2(0,T_0;L^2(\omega)).$$

From standard results on diffusion convection equations (on continuity of the solution with respect to the right-hand side, the initial data φ_0 being fixed) and compactness embeddings, we then have

$$\varphi_\beta \to \tilde{\varphi} \text{ in } C([0,T_0];L^2(\Omega)),$$

where $\tilde{\varphi}$ denotes the solution of (2.31), (2.32), (2.33), and (2.34) corresponding to $\tilde{v}$. Therefore $\varphi_\beta(0) \to \tilde{\varphi}(0)$ in $L^2(\Omega)$ and we must have

$$\tilde{\varphi}(0) = 0.$$

Because of the definition of $\bar{v}$ we then have

$$\int_0^{T_0}\int_\omega |\bar{v}|^2 dxdt \le \int_0^{T_0}\int_\omega |\tilde{v}|^2 dxdt.$$

But on the other hand, because of the weak convergence of v_β to $\tilde{v}$, we have

$$\int_0^{T_0}\int_\omega |\tilde{v}|^2 dxdt \le \int_0^{T_0}\int_\omega |\bar{v}|^2 dxdt$$

so that the convergence of v_β to $\tilde{v}$ is strong and $\tilde{v}$ minimizes the norm among the elements v such that (2.31), (2.32), (2.33), and (2.34) are satisfied. By uniqueness of this minimum, we must have $\tilde{v} = \bar{v}$. Now, in view of the previous results, (2.35) is clear and the proof of Theorem 2.12 is complete. □

Without any further assumption, it seems impossible to obtain a rate of convergence in the previous approximation. Nevertheless, we will obtain a rate of convergence under a regularity assumption on the "adjoint state" associated to the null controllability problem solved in Theorem 2.9. We first have to introduce this adjoint state.

We recall that

$$(p,q) \in V^0 \to (p,q)_{V^0} = \int_0^{T_0}\int_\omega pq dxdt$$

is the scalar product on V^0. On the other hand, the mapping

$$q \to \int_\Omega \varphi_0 q(T_0) dx$$

is a linear continuous form on V^0. Therefore, from the Riesz theorem, there exists a unique $p \in V^0$ such that

$$\forall q \in V^0,\ \int_0^{T_0}\int_\omega pq dxdt = \int_\Omega \varphi_0 q(T_0) dx. \tag{2.36}$$

We now use the results of Theorem 2.12. The adjoint state p_β corresponding to the approximate optimal control problem is an element of V^0. Moreover, because $p_\beta = -v_\beta$ and v_β converges strongly to $\bar{v}$ in $L^2(0,T_0;L^2(\omega))$, we see that p_β converges strongly to $\bar{p}$ in V^0 with

$$\bar{p} = \bar{v} \text{ in } \omega \times (0,T_0).$$

Now let us multiply by q, where $q \in \mathcal{V}^0$, the equation satisfied by φ_β. Integrating by parts and using the properties of q, we obtain

$$\int_0^{T_0}\int_\omega v_\beta q dxdt = -\int_\Omega \varphi_\beta(T_0)q(T_0)dx + \int_\Omega \varphi_\beta(0)q(0)dx.$$

Therefore

$$\int_0^{T_0}\int_\omega p_\beta q dxdt = \int_\Omega \varphi_0 q(T_0)dx - \int_\Omega \varphi_\beta(0)q(0)dx.$$

We can now pass to the limit when $\beta \to 0$ using the convergence of $\varphi_\beta(0)$ to 0 and obtain

$$\forall q \in \mathcal{V}^0, \quad \int_0^{T_0}\int_\omega \bar{p} q dxdt = \int_\Omega \varphi_0 q(T_0)dx.$$

As $\mathcal{V}^0$ is dense in V^0, this is also valid for every $q \in V^0$, which says that $\bar{p}$ is a solution of (2.36). By uniqueness of this solution we have $\bar{p} = p$.

Notice that because of the definition of V^0, the (uniquely defined) function $\bar{p}$ may not have any value at $t = 0$ a priori, but satisfies the following system:

$$\frac{\partial \bar{p}}{\partial t} - \sum_{i,j=1}^N \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial \bar{p}}{\partial x_j}\right) + \sum_{i=1}^N b_i \frac{\partial \bar{p}}{\partial x_i} + \sum_{j=1}^N \frac{\partial}{\partial x_j}(c_j \bar{p}) = 0 \text{ in } \Omega \times (0,T_0), \tag{2.37}$$

$$\bar{p} = 0 \text{ on } \Gamma \times (0,T_0), \tag{2.38}$$

$$\bar{p} + \bar{v} = 0 \text{ in } \omega \times (0,T_0). \tag{2.39}$$

We are now ready to give a result concerning the convergence rates of the approximation under a regularity assumption on the adjoint state $\bar{p}$.

THEOREM 2.13. *We use the notation of Theorem* 2.12. *Let us assume that the function $\bar{p}$, solution of* (2.36), *satisfies*

$$\bar{p} \in C([0,T_0];L^2(\Omega)). \tag{2.40}$$

Then we have

$$|\varphi_\beta(0)|_{L^2(\Omega)} \leq 2\beta|\bar{p}(0)|_{L^2(\Omega)}, \tag{2.41}$$

$$|v_\beta - \bar{v}|_{L^2(0,T_0;L^2(\omega))} \leq 2\beta^{\frac{1}{2}}|\bar{p}(0)|_{L^2(\Omega)}, \tag{2.42}$$

$$\left|\int_\Omega y(T_0)\varphi_0 dx - \int_0^{T_0}\int_\Omega f\varphi_\beta dxdt + \int_0^{T_0}\int_\omega h.v_\beta dxdt\right| \leq C\beta^{\frac{1}{2}}|\bar{p}(0)|_{L^2(\Omega)}, \tag{2.43}$$

where C is independent of β and of φ_0.

Proof. We can write

$$\begin{aligned}
&\frac{1}{2}\int_0^{T_0}\int_\omega |v_\beta-\bar{v}|^2dxdt+\frac{1}{2\beta}\int_\Omega|\varphi_\beta(0)|^2dx\\
&=\frac{1}{2}\int_0^{T_0}\int_\omega |v_\beta|^2dxdt+\frac{1}{2\beta}\int_\Omega|\varphi_\beta(0)|^2dx+\frac{1}{2}\int_0^{T_0}\int_\omega|\bar{v}|^2dxdt-\int_0^{T_0}\int_\omega v_\beta\bar{v}dxdt\\
&=K_\beta(v_\beta)+\frac{1}{2}\int_0^{T_0}\int_\omega|\bar{v}|^2dxdt-\int_0^{T_0}\int_\omega v_\beta\bar{v}dxdt\\
&\le K_\beta(\bar{v})+\frac{1}{2}\int_0^{T_0}\int_\omega|\bar{v}|^2dxdt-\int_0^{T_0}\int_\omega v_\beta\bar{v}dxdt\\
&=\int_0^{T_0}\int_\omega|\bar{v}|^2dxdt-\int_0^{T_0}\int_\omega v_\beta\bar{v}dxdt=\int_0^{T_0}\int_\omega(\bar{v}-v_\beta)\bar{v}dxdt.
\end{aligned}$$

The function $\phi=(\bar{\varphi}-\varphi_\beta)$ satisfies the following system:

$$\begin{aligned}
&-\frac{\partial\phi}{\partial t}-\sum_{i,j=1}^N\frac{\partial}{\partial x_j}\left(a_{ji}\frac{\partial\phi}{\partial x_i}\right)-\sum_{i=1}^N\frac{\partial}{\partial x_i}(b_i\phi)-\sum_{j=1}^N c_j\frac{\partial\phi}{\partial x_j}=(\bar{v}-v_\beta)\chi_\omega \text{ in } \Omega\times(0,T_0),\\
&\phi=0 \text{ on } \Gamma\times(0,T_0),\\
&\phi(T_0)=0 \text{ in } \Omega,\\
&\phi(0)=-\varphi_\beta(0) \text{ in } \Omega.
\end{aligned}$$

Multiplying this equation by $\bar{p}$, using (2.39), and integrating by parts, we obtain

$$\int_0^{T_0}\int_\omega(\bar{v}-v_\beta)\bar{v}dxdt=\int_\Omega\varphi_\beta(0)\bar{p}(0)dx.$$

Therefore we have

$$\frac{1}{2}\int_0^{T_0}\int_\omega|v_\beta-\bar{v}|^2dxdt+\frac{1}{2\beta}\int_\Omega|\varphi_\beta(0)|^2dx\le|\varphi_\beta(0)|_{L^2(\Omega)}|\bar{p}(0)|_{L^2(\Omega)}.$$

This gives immediately (2.41) and (2.42). The estimate on $(\bar{\varphi}-\varphi_\beta)$ is of the same order as the one on $(\bar{v}-v_\beta)$, and this implies (2.43) which finishes the proof of Theorem 2.13. □

Remark 2.14. For each φ_0, in order to find an approximation of $\int_\Omega y(T_0)\varphi_0dx$, we have to solve a classical optimal control problem for the adjoint system. We have to notice that for different elements φ_0 the optimal control problems to be solved are essentially the same and differ only in the initial data in (2.31). This is particularly important for numerical approximation because all the linear systems corresponding to different elements φ_0 have the same matrices.

2.2. Linearized Navier–Stokes equations. We consider here in dimension $N=3$ the Navier–Stokes equations linearized around a velocity $\bar{y}$ such that

$$\bar{y}\in L^\infty(0,T_0;W^{1,\infty}(\Omega)),\ \frac{\partial\bar{y}}{\partial t}\in L^2(0,T_0;W^{1,\sigma}(\Omega)),\ \sigma>\frac{6}{5},\ \operatorname{div}(\bar{y})=0, \tag{2.44}$$

namely,

$$\frac{\partial y}{\partial t}-\mu\Delta y+(\bar{y}.\nabla)y+(y.\nabla)\bar{y}+\nabla p=f \text{ in } \Omega\times(0,T_0), \tag{2.45}$$

$$\operatorname{div} y=0 \text{ in } \Omega\times(0,T_0), \tag{2.46}$$

$$y=0 \text{ on } \Gamma\times(0,T_0), \tag{2.47}$$

where $f\in L^2(0,T_0;(L^2(\Omega))^3)$ and $\mu>0$.

Here again we either do not impose any initial condition on y or do not know $y(0)$. We suppose that we know a measurement of the solution on a subdomain

$$y_{/\omega\times(0,T_0)} = h, \tag{2.48}$$

where ω is a nonempty open set contained in Ω, and $h \in L^2(0,T_0;(L^2(\omega))^3)$.

Remark 2.15. (1) We could also consider the case where normal stresses

$$(\sigma.\nu)_i = -p\nu_i + \mu\sum_{j=1}^{3} D_{ij}(y)\nu_j,\ i = 1,\dots,3,$$

with $D_{ij}(y) = \frac{1}{2}(\frac{\partial y_i}{\partial x_j} + \frac{\partial y_j}{\partial x_i})$ are known on $\Gamma_0 \times (0,T_0)$, where Γ_0 is a nonempty relatively open set of the boundary Γ.

(2) We have taken here the case of measurements on y only, but we could also have local measurements on the pressure p. This would correspond to a simpler situation. On the other hand, the case of measurements only on the pressure turns out to be impossible to treat.

In the case of classical variational data assimilation we take the initial value as a control variable

$$y(0) = y_0 \text{ in } \Omega, \tag{2.49}$$

where

$$y_0 \in H = \{z \in (L^2(\Omega))^3,\ \operatorname{div} z = 0,\ z.\nu = 0 \text{ on } \Gamma\}. \tag{2.50}$$

We know (cf. [18] or [22]) that for every $y_0 \in H$, there exists a unique solution $y = y[y_0]$ of (2.45), (2.46), (2.47), (2.49) with

$$y[y_0] \in C([0,T_0];H) \cap L^2(0,T_0;(H_0^1(\Omega))^3).$$

We then want to find y_0 such that the error between the actual measure h and the value of the solution $y = y[y_0]$ of (2.45), (2.46), (2.47), (2.49) on the subdomain $\omega \times (0,T_0)$ achieves its minimum. If we define

$$\mathbf{J}(y_0) = \frac{1}{2}\int_0^{T_0}\int_\omega |y[y_0] - h|^2 dxdt, \tag{2.51}$$

we consider the following optimal control problem:

$$\begin{aligned}&\text{Find } \bar{y}_0 \in H \text{ such that}\\ &\mathbf{J}(\bar{y}_0) = \min_{y_0\in H} \mathbf{J}(y_0).\end{aligned} \tag{2.52}$$

Again this problem is ill-posed and we must add a Tychonov regularization term by considering for $\alpha > 0$,

$$\mathbf{J}_\alpha(y_0) = \frac{1}{2}\int_0^{T_0}\int_\omega |y[y_0] - h|^2 dxdt + \frac{\alpha}{2}|y_0|_H^2. \tag{2.53}$$

We now solve the following regularized optimal control problem:

$$\begin{aligned}&\text{Find } y_\alpha \in H \text{ such that}\\ &\mathbf{J}_\alpha(y_\alpha) = \min_{y_0\in H} \mathbf{J}_\alpha(y_0).\end{aligned} \tag{2.54}$$

This problem is classical and has a unique solution $y_\alpha \in H$, and the questions are again to give a meaning to this solution, to understand what happens when $\alpha \to 0$, and to estimate the sensitivity to errors in the measurements h.

Following the same ideas as in the previous section, we will present a nonstandard approach using controllability techniques.

Let us define

$$\mathcal{W} = \{y,\ y \text{ solution of (2.45), (2.46), (2.47), (2.49)},\ y_0 \in H\}.$$

If we call y^0 the element of $\mathcal{W}$ such that $y_0 = 0$, we have

$$\mathcal{W} = y^0 + \mathcal{W}^0,$$

where $\mathcal{W}^0$ is a vector space.

We want to give a Carleman estimate for elements of $\mathcal{W}$. We use here the results of [12], [13], and [7] concerning Navier–Stokes equations, and the Carleman estimate in this case is much more difficult to obtain than for diffusion convection equations. We take the same principal weight ψ as in the previous section, but this time we need to take functions ξ_4 and η_4 defined in (2.17) and (2.18). Moreover, we must define

$$\begin{aligned} \widehat{\xi}_4(t) &= \max_{x\in\overline{\Omega}} \xi_4(x,t), \\ \widehat{\eta}_4(t) &= \min_{x\in\overline{\Omega}} \eta_4(x,t), \\ \eta_4^*(t) &= \max_{x\in\overline{\Omega}} \eta_4(x,t), \\ \widehat{\theta}(t) = s\lambda e^{-s\widehat{\eta}_4}\widehat{\xi}_4, &\quad \theta(t) = s^{15/4} e^{-2s\widehat{\eta}_4 + s\eta_4^*}\widehat{\xi}_4^{15/4}. \end{aligned} \tag{2.55}$$

We then have the following.

PROPOSITION 2.16. *There exist $s_1 > 0$ and $\lambda_1 > 0$, and there exists a constant C depending on Ω, ω, ψ, μ, T_0, and $\bar{y}$ (in the spaces corresponding to hypothesis* (2.44)) *such that for every $s > s_1$, for every $\lambda > \lambda_1$, and for every $y \in \mathcal{W}$, we have*

$$\begin{aligned} s^3\lambda^4 \int_0^{T_0}\int_\Omega e^{-2s\eta_4}\xi_4^3|y|^2\,dx\,dt + s\lambda^2\int_0^{T_0}\int_\Omega e^{-2s\eta_4}\xi_4|\nabla y|^2\,dx\,dt \\ \le C\left(s^{15/2}\lambda^{20}\int_0^{T_0}\int_\Omega e^{-4s\widehat{\eta}_4+2s\eta^*}\widehat{\xi}_4^{15/2}|f|^2\,dx\,dt \right. \\ \left. + s^{16}\lambda^{40}\int_0^{T_0}\int_\omega e^{-8s\widehat{\eta}_4+6s\eta^*}\widehat{\xi}_4^{16}|y|^2\,dx\,dt\right). \end{aligned} \tag{2.56}$$

From now on, we fix $s > s_1$ and $\lambda > \lambda_1$ and repeat the arguments of Proposition 2.4. Using (2.56) we can show the following unique continuation property:

$$\forall z \in \mathcal{W}^0, \quad z = 0 \text{ on } \omega \times (0,T_0) \ \Rightarrow\ z = 0 \text{ in } \Omega \times (0,T_0).$$

Therefore the bilinear form defined by

$$\forall z, \tilde{z} \in \mathcal{W}^0,\ (z,\tilde{z})_{W^0} = \int_0^{T_0}\int_\omega z.\tilde{z}dxdt$$

is a scalar product on $\mathcal{W}^0$ and we can set the following definition.

DEFINITION 2.17. *We denote by W^0 the (abstract) completion of $\mathcal{W}^0$ with respect to the norm $|.|_{W^0}$ associated with the above defined scalar product $(.,.)_{W^0}$ and denote by W the translated space*

$$W = y^0 + W^0. \tag{2.57}$$

Then W^0 is a Hilbert space for the scalar product $(.,.)_{W^0}$, and for every $y \in W$, inequality (2.56) still holds true.

Remark 2.18. Because of inequality (2.56) and classical energy estimates for the linearized Navier–Stokes operator, we can easily show that for every $\delta > 0$, any element y of W satisfies $y \in C([\delta, T_0]; H) \cap L^2(\delta, T_0; (H_0^1(\Omega))^3)$. In particular, the value $y(T_0)$ makes perfect sense in H. But, a priori, the function y may have no value at $t = 0$ (initial value) in any sense.

We now consider the backward adjoint controlled problem for a distributed control $v \in L^2(0, T_0; (L^2(\omega))^3)$,

$$-\frac{\partial \varphi}{\partial t} - \mu\Delta\varphi - (\bar{y}.\nabla)\varphi + (\nabla\bar{y})\varphi + \nabla\pi = v.\chi_\omega \text{ in } \Omega \times (0, T_0), \tag{2.58}$$

$$\operatorname{div} \varphi = 0 \text{ in } \Omega \times (0, T_0), \tag{2.59}$$

$$\varphi = 0 \text{ on } \Gamma \times (0, T_0), \tag{2.60}$$

$$\varphi(T_0) = \varphi_0, \tag{2.61}$$

where $\varphi_0 \in H$ and H is defined in (2.50).

Using the same arguments as in Theorem 2.9 we obtain the following result of stability and reconstruction of $y(T_0)$.

THEOREM 2.19. *Under the previous hypotheses, for every $\omega \subset \Omega$, for every $T_0 > 0$, and for every $\varphi_0 \in H$, there exists $v = v(\varphi_0) \in L^2(0, T_0; (L^2(\omega))^3)$ such that the solution φ of* (2.58), (2.59), (2.60), (2.61) *satisfies*

$$\varphi(0) = 0. \tag{2.62}$$

We will choose $v(\varphi_0)$ of minimal norm among the controls such that (2.62) *is satisfied, and then the mapping $\varphi_0 \to v(\varphi_0)$ is continuous, which says that*

$$\exists C > 0 \ \forall \varphi_0 \in H, \ |v(\varphi_0)|_{L^2(0,T_0;(L^2(\omega))^3)} \leq C|\varphi_0|_H. \tag{2.63}$$

We then have, if $y = h$ on $\omega \times (0, T_0)$,

$$\forall \varphi_0 \in H, \ (y(T_0), \varphi_0)_H = \int_0^{T_0} \int_\Omega f.\varphi dxdt - \int_0^{T_0} \int_\omega h.v(\varphi_0) dxdt. \tag{2.64}$$

Moreover, there exists a constant $C > 0$ depending only on Ω, ω, T_0, μ, and $\bar{y}$ (*in the spaces corresponding to hypothesis* (2.44)) *such that*

$$|y(T_0)|_H^2 \leq C \left(\int_0^{T_0} \int_\Omega |f|^2 dxdt + \int_0^{T_0} \int_\omega |h|^2 dxdt \right). \tag{2.65}$$

We also obtain, in the same way as in Corollary 2.11, a result measuring the sensitivity of the recovered state with respect to errors in the measurements.

COROLLARY 2.20. *We write $y(T_0)$ for the recovery obtained by the previous method from a measurement h and write $\hat{y}(T_0)$ for the recovery obtained using a measurement $\hat{h}$. Then there exists a constant $C > 0$, independent of h and $\hat{h}$, such that*

$$|y(T_0) - \hat{y}(T_0)|_H \leq C|h - \hat{h}|_{L^2(0,T_0;(L^2(\omega))^3)}. \tag{2.66}$$

Here also we can consider an optimal control problem which will provide an approximation for $(y(T_0), \varphi_0)_H$. For $v \in L^2(0, T_0; (L^2(\omega))^3)$ let φ be solution of (2.58), (2.59), (2.60), (2.61), and for $\beta > 0$ let us define a cost function $\mathbf{K}_\beta(v)$ by

$$\mathbf{K}_\beta(v) = \frac{1}{2\beta}|\varphi(0)|_H^2 + \frac{1}{2}\int_0^{T_0}\int_\omega |v|^2 dxdt. \tag{2.67}$$

We look for $v_\beta \in L^2(0, T_0; (L^2(\omega))^3)$ such that

$$\mathbf{J}_\beta(v_\beta) = \min_{v\in L^2(0,T_0;(L^2(\omega))^3)} \mathbf{J}_\beta(v). \tag{2.68}$$

For fixed $\beta > 0$ this last problem is a classical optimal control problem. We then obtain, following the same argument as for Theorem 2.12, the following theorem.

THEOREM 2.21. (1) *For every $\beta > 0$, there exists a unique solution v_β to* (2.68), *and v_β is characterized by the following optimality system:*

$$\begin{aligned}
&-\frac{\partial\varphi_\beta}{\partial t} - \mu\Delta\varphi_\beta - (\bar{y}.\nabla)\varphi_\beta + (\nabla\bar{y})\varphi_\beta + \nabla\pi = v_\beta.\chi_\omega \ \textit{in}\ \Omega \times (0, T_0),\\
&\operatorname{div}\varphi_\beta = 0 \ \textit{in}\ \Omega \times (0, T_0),\\
&\varphi_\beta = 0 \ \textit{on}\ \Gamma \times (0, T_0),\\
&\varphi_\beta(T_0) = \varphi_0,\\
&\frac{\partial r_\beta}{\partial t} - \mu\Delta r_\beta + (\bar{y}.\nabla)r_\beta + (r_\beta.\nabla)\bar{y} + \nabla\rho = 0 \ \textit{in}\ \Omega \times (0, T_0),\\
&\operatorname{div} r_\beta = 0 \ \textit{in}\ \Omega \times (0, T_0),\\
&r_\beta = 0 \ \textit{on}\ \Gamma \times (0, T_0),\\
&r_\beta(0) = \frac{1}{\beta}\varphi_\beta(0),\\
&r_\beta + v_\beta = 0 \ \textit{in}\ \omega \times (0, T_0).
\end{aligned}$$

(2) *When β tends to zero,*

$$v_\beta \to \bar{v} \ \textit{in}\ L^2(0, T_0; (L^2(\omega))^3), \ \varphi_\beta \to \bar{\varphi} \ \textit{in}\ C([0, T_0]; H),$$

where $\bar{\varphi}$ and $\bar{v}$ satisfy (2.58)–(2.61) *and* (2.62). *Moreover, $\bar{v}$ is the element with minimal norm such that* (2.58)–(2.61) *and* (2.62) *are satisfied. In addition, when, $\beta \to 0$, we have*

$$\int_0^{T_0}\int_\Omega f.\varphi_\beta dxdt - \int_0^{T_0}\int_\omega y.v_\beta dxdt \to (y(T_0), \varphi_0)_H. \tag{2.69}$$

We can obtain an adjoint state corresponding to the null controllability problem solved in Theorem 2.19. The mapping $q \to (\varphi_0, q(T_0))_H$ is a continuous linear form on W^0. Therefore, from the Riesz theorem, there exists a unique function $\bar{r} \in W^0$ such that

$$\forall q \in W^0, \ (\bar{r}, q)_{W^0} = (\varphi_0, q(T_0))_H. \tag{2.70}$$

It is easy to show, as in the case of diffusion convection equations, that r_β converges to $\bar{r}$ in W^0 and that $\bar{r}+\bar{v}=0$ on $\omega\times(0,T_0)$. We only know a priori that $\bar{r}$ is an element of W^0 so that it may not have any initial value at $t=0$. But under an additional regularity assumption on $\bar{r}$ we can obtain an estimate on the convergence rate of the previous approximation procedure.

THEOREM 2.22. *We use the notation of Theorem* 2.21. *Let us assume that the function $\bar{r}$, solution of* (2.70)*, satisfies*

$$\bar{r}\in C([0,T_0];H). \tag{2.71}$$

Then we have

$$|\varphi_\beta(0)|_H\le 2\beta|\bar{r}(0)|_H, \tag{2.72}$$

$$|v_\beta-\bar{v}|_{L^2(0,T_0;(L^2(\omega))^3)}\le 2\beta^{\frac12}|\bar{r}(0)|_H, \tag{2.73}$$

$$\left|(y(T_0),\varphi_0)_H-\int_0^{T_0}\int_\Omega f.\varphi_\beta dxdt+\int_0^{T_0}\int_\omega h.v_\beta dxdt\right|\le C\beta^{\frac12}|\bar{r}(0)|_H, \tag{2.74}$$

where C is independent of β and of φ_0.

The proof uses exactly the same arguments as those used for the proof of Theorem 2.13.

3. New results on Tychonov regularization. We are going to give an existence result for a nonclassical optimal control problem, and this will enable us to explain the reason for which in practical situations, with natural hypotheses, Tychonov regularization works in the sense that the corresponding solution converges when the regularization parameter α tends to zero.

Let us for the moment go back to the case of diffusion convection equations (we keep the same notation), as the other cases can be treated exactly in the same way.

We recall that a function y in V is a solution of (2.1) and (2.2). But y may not have a value at $t=0$ (which will be referred to as an initial value) in any sense. Let us define for $y\in V$

$$\tilde{J}(y)=\int_0^{T_0}\int_\omega |y-h|^2dxdt. \tag{3.1}$$

Of course we notice that if y has an initial value y_0, then $\tilde{J}(y)=J(y_0)$, but it is no longer the case in general. It is essential to understand that we consider the same functional value but defined on a different argument. Notice also that $\tilde{J}$ is perfectly defined for $y\in V$, as we know that for $y\in V$, we have $\int_0^{T_0}\int_\omega|y|^2dxdt<+\infty$. It is now immediate to obtain the following result.

THEOREM 3.1. *There exists a unique element $\tilde{y}\in V$, which is a solution of the following minimization problem:*

$$\begin{aligned}&\textit{Find } \tilde{y}\in V \textit{ such that}\\ &\tilde{J}(\tilde{y})=\min_{y\in V}\tilde{J}(y).\end{aligned} \tag{3.2}$$

Moreover if $h,h^0\in L^2(0,T_0;L^2(\omega))$ and the corresponding solutions are called $\tilde{y}$ and $\tilde{y}^0$, then we have

$$|\tilde{y}-\tilde{y}^0|^2_{V^0}\le\int_0^{T_0}\int_\omega|h-h^0|^2dxdt, \tag{3.3}$$

which implies that for every $\delta > 0$, there exists a constant $C(\delta) > 0$ such that

$$|\tilde{y} - \tilde{y}^0|^2_{C([\delta,T_0];L^2(\Omega))} + ||\tilde{y} - \tilde{y}^0||^2_{L^2(\delta,T_0;H^1_0(\Omega))} \leq C(\delta) \int_0^{T_0} \int_\omega |h - h^0|^2 dxdt.$$

Proof. The proof of existence and of the stability inequality is elementary, as V is a Hilbert space for the norm

$$y \to \left(\int_0^{T_0} \int_\omega |y|^2 dxdt \right)^{\frac{1}{2}}.$$

Uniqueness also follows immediately from the unique continuation property, which is valid on V^0. The last estimate comes from (2.23).

Of course the previous result is immediate once we know all the technicalities which are included in the Carleman estimates.

Remark 3.2. Therefore, a problem which was ill-posed if we minimize with respect to the initial value has become well-posed when minimizing with respect to the trajectory in V. We also have an estimate of the sensitivity to errors in the measurements h. This estimate is given in the V-norm, but thanks to (2.21) and (2.23) it is also valid in the weighted Sobolev spaces which occur in Carleman estimates and in classical Sobolev spaces away from $t = 0$.

Now let us make a regularity hypothesis on the minimizer $\tilde{y}$, namely, that it has an initial value in the sense that

$$\tilde{y} \in C([0, T_0]; L^2(\Omega)), \ \tilde{y}(0) = \tilde{y}_0 \text{ in } \Omega. \tag{3.4}$$

This regularity hypothesis is quite natural in practical (realistic) situations, even if this depends strongly on the (real) measurements which we are dealing with and which are not usually the true value of a solution on the subdomain $\omega \times (0, T_0)$.

Then we have the following convergence theorem.

THEOREM 3.3. *Let us suppose that* (3.4) *is true. Then we have*

$$J(\tilde{y}_0) = \tilde{J}(\tilde{y}),$$

and $\tilde{y}_0$ is solution to the optimal control problem (2.9).

We also have the following estimate:

$$\int_0^{T_0} \int_\omega |\tilde{y} - y^\alpha|^2 dxdt \leq \alpha |\tilde{y}_0|_{L^2(\Omega)}. \tag{3.5}$$

Moreover, when $\alpha \to 0$, the solution y_α of the Tychonov regularized optimal control problem converges strongly in $L^2(\Omega)$ to $\tilde{y}_0$.

Proof. Let us call y^α the solution of (2.1), (2.2), and (2.7) with $y^\alpha(0) = y_\alpha$. We have

$$J_\alpha(y_\alpha) \leq J_\alpha(\tilde{y}_0)$$

and

$$\tilde{J}(\tilde{y}) \leq \tilde{J}(y^\alpha).$$

Therefore

$$\begin{aligned}\frac{1}{2}\int_0^{T_0}\int_\omega |y^\alpha - h|^2 dxdt + \frac{\alpha}{2}|y_\alpha|^2_{L^2(\Omega)} &\leq \frac{1}{2}\int_0^{T_0}\int_\omega |\tilde{y} - h|^2 dxdt + \frac{\alpha}{2}|\tilde{y}_0|^2_{L^2(\Omega)}\\ &\leq \frac{1}{2}\int_0^{T_0}\int_\omega |y^\alpha - h|^2 dxdt + \frac{\alpha}{2}|\tilde{y}_0|^2_{L^2(\Omega)}.\end{aligned}$$

As a consequence we have

$$\forall \alpha > 0,\ |y_\alpha|^2_{L^2(\Omega)} \leq |\tilde{y}_0|^2_{L^2(\Omega)}.$$

After extraction of a subsequence (still denoted by y_α), we can suppose that

$$y_\alpha \rightharpoonup \hat{y}_0 \text{ in } L^2(\Omega) \text{ weakly},$$

and if $\hat{y}$ denotes the solution of (2.1), (2.2), and (2.7) with $\hat{y}(0) = \hat{y}_0$, we have

$$y^\alpha \to \hat{y} \text{ in } L^2(0, T_0; L^2(\Omega))$$

and also in various topologies.

Now we have

$$\frac{\alpha}{2}|y_\alpha|^2_{L^2(\Omega)} \to 0 \text{ when } \ \alpha \to 0$$

so that necessarily, when $\alpha \to 0$,

$$\frac{1}{2}\int_0^{T_0}\int_\omega |y^\alpha - h|^2 dxdt \to \frac{1}{2}\int_0^{T_0}\int_\omega |\tilde{y} - h|^2 dxdt.$$

Therefore we must have

$$\frac{1}{2}\int_0^{T_0}\int_\omega |\hat{y} - h|^2 dxdt = \frac{1}{2}\int_0^{T_0}\int_\omega |\tilde{y} - h|^2 dxdt.$$

From uniqueness in problem (3.2), we see that necessarily we have

$$\hat{y} = \tilde{y} \text{ in } \Omega \times (0, T_0),$$

so that

$$\hat{y}_0 = \tilde{y}_0.$$

We now know that

$$y_\alpha \rightharpoonup \tilde{y}_0 \text{ in } L^2(\Omega) \text{ weakly}$$

and

$$|y_\alpha|^2_{L^2(\Omega)} \leq |\tilde{y}_0|^2_{L^2(\Omega)}.$$

This implies strong convergence in $L^2(\Omega)$ of y_α towards $\tilde{y}_0$.

As $\tilde{y}$ is the minimizer of $\tilde{J}$, we see from the Euler–Lagrange equation associated to this minimization problem that

$$\forall z \in V,\ \int_0^{T_0}\int_\omega (\tilde{y} - h).(\tilde{y} - z)dxdt = 0. \tag{3.6}$$

Now a simple calculation gives

$$\begin{aligned}&\int_0^{T_0}\int_\omega |\tilde{y}-y^\alpha|^2 dxdt + \alpha|\tilde{y}_0 - y_\alpha|^2_{L^2(\Omega)}\\ &\quad = \int_0^{T_0}\int_\omega |y^\alpha - h|^2 dxdt + \alpha|y_\alpha|^2_{L^2(\Omega)} + \int_0^{T_0}\int_\omega |\tilde{y}-h|^2 dxdt\\ &\qquad + \alpha|\tilde{y}_0|^2_{L^2(\Omega)} - 2\int_0^{T_0}\int_\omega (\tilde{y}-h).(y^\alpha - h)dxdt - 2\alpha(\tilde{y}_0, y_\alpha)_{L^2(\Omega)}\\ &\quad \le 2\int_0^{T_0}\int_\omega |\tilde{y}-h|^2 dxdt - 2\int_0^{T_0}\int_\omega (\tilde{y}-h).(y^\alpha - h)dxdt\\ &\qquad + 2\alpha|\tilde{y}_0|^2_{L^2(\Omega)} - 2\alpha(\tilde{y}_0, y_\alpha)_{L^2(\Omega)}\\ &\quad \le 2\int_0^{T_0}\int_\omega (\tilde{y}-h).(\tilde{y}-y^\alpha)dxdt + 2\alpha(\tilde{y}_0, \tilde{y}_0 - y_\alpha)_{L^2(\Omega)}.\end{aligned}$$

But as $y^\alpha \in V$, because of (3.6) we obtain

$$\int_0^{T_0}\int_\omega |\tilde{y}-y^\alpha|^2 dxdt + \alpha|\tilde{y}_0 - y_\alpha|^2_{L^2(\Omega)} \le 2\alpha(\tilde{y}_0, \tilde{y}_0 - y_\alpha)_{L^2(\Omega)}. \tag{3.7}$$

This gives immediately (3.5), and the proof of Theorem 3.3 is now complete. □

Remark 3.4. (1) Our situation is quite different from the one which is considered classically, for example, in [10], [5], or [4]. These authors make an a priori assumption, which is usually written in an abstract form, which essentially assumes that there exists a solution to the minimization problem for the functional J. Here, we prove the existence and uniqueness of a minimizer for $\tilde{J}$ (without an additional hypothesis) and, afterwards, we make a regularity assumption on this minimizer.

(2) We obtain a rate of convergence in the V-distance but, without additional hypotheses, this does not give any rate of convergence for the initial values $(\tilde{y}_0 - y_\alpha)$.

(3) In the same way, the last theorem does not say anything about the behavior of solutions y_α with very small α when we have perturbations in the measurements h. Even if the hypothesis (3.4) seems "natural" in practical situations, it is really unnatural to assume that this initial value, that we assume to exist, would be continuous with respect to h. This is exactly what is missing in the information given by Carleman estimates.

In order to give an estimate for the rate of convergence for the initial values in the Tychonov regularization method, we will follow a method similar to the one used in [10], [5], and [4], but with a different hypothesis. We will make an assumption on $\tilde{y}_0$, which will be made precise below, but which, roughly speaking, says that $\tilde{y}_0$ is on a *controlled* trajectory of the adjoint operator with control acting everywhere in the domain, except in the neighborhood of $t = 0$ where it can act only on ω.

THEOREM 3.5. *Let us assume, in addition to* (3.4)*, that there exist* $q_0 \in L^2(\Omega)$*,* $w \in L^2(0, T_0; L^2(\omega))$*, and* g *such that*

$$\frac{e^{s\tilde{\eta}_1}}{\tilde{\xi}_1^{\frac{3}{2}}} g \in L^2(0, T_0; L^2(\Omega))$$

(this will be the case if, for example, $g \in L^2(0, T_0; L^2(\Omega))$*) and* $g = 0$ *in a neighborhood*

of $t = 0$) such that the solution q of the equation

$$-\frac{\partial q}{\partial t} - \sum_{i,j=1}^{N} \frac{\partial}{\partial x_j}\left(a_{ji}\frac{\partial q}{\partial x_i}\right) - \sum_{i=1}^{N} \frac{\partial}{\partial x_i}(b_i q) - \sum_{j=1}^{N} c_j \frac{\partial q}{\partial x_j} = g + w.\chi_\omega \quad \textit{in } \Omega \times (0, T_0), \tag{3.8}$$

$$q = 0 \textit{ on } \Gamma \times (0, T_0), \tag{3.9}$$

$$q(T_0) = q_0 \textit{ in } \Omega \tag{3.10}$$

satisfies

$$q(0) = \tilde{y}_0. \tag{3.11}$$

Then there exists a constant $C > 0$ depending on g, w, and q_0 such that

$$\int_0^{T_0} \int_\omega |\tilde{y} - y^\alpha|^2 dxdt \leq C\alpha^2, \tag{3.12}$$

$$|\tilde{y}_0 - y_\alpha|^2_{L^2(\Omega)} \leq C\alpha. \tag{3.13}$$

Remark 3.6. Due to the null controllability property for the adjoint operator with control acting only on ω, it can be shown that the hypothesis made in Theorem 3.5 is equivalent to the same hypothesis with $g = 0$ and $q_0 = 0$.

Proof. From (3.7) we know that

$$\int_0^{T_0} \int_\omega |\tilde{y} - y^\alpha|^2 dxdt + \alpha|\tilde{y}_0 - y_\alpha|^2_{L^2(\Omega)} \leq 2\alpha(\tilde{y}_0, \tilde{y}_0 - y_\alpha)_{L^2(\Omega)}.$$

Let us multiply (3.8) by $(\tilde{y} - y^\alpha)$ and integrate by parts. We obtain, using the equation satisfied by $(\tilde{y} - y^\alpha)$,

$$\begin{aligned}(\tilde{y}_0, \tilde{y}_0 - y_\alpha)_{L^2(\Omega)} &= (q_0, \tilde{y}(T_0) - y^\alpha(T_0))_{L^2(\Omega)} + \int_0^{T_0}\int_\Omega g(\tilde{y} - y^\alpha)dxdt \\ &\quad + \int_0^{T_0}\int_\omega w(\tilde{y} - y^\alpha)dxdt \leq |q_0|_{L^2(\Omega)}|\tilde{y}(T_0) - y^\alpha(T_0)|_{L^2(\Omega)} \\ &\quad + \left(\int_0^{T_0}\int_\Omega \frac{e^{2s\tilde{\eta}_1}}{\tilde{\xi}_1^3}|g|^2 dxdt\right)^{\frac{1}{2}} \left(\int_0^{T_0}\int_\Omega \tilde{\xi}_1^3 e^{-2s\tilde{\eta}_1}|\tilde{y} - y^\alpha|^2 dxdt\right)^{\frac{1}{2}} \\ &\quad + \left(\int_0^{T_0}\int_\omega |w|^2 dxdt\right)^{\frac{1}{2}} \left(\int_0^{T_0}\int_\omega |\tilde{y} - y^\alpha|^2 dxdt\right)^{\frac{1}{2}}.\end{aligned}$$

But from the Carleman estimate (2.21) and the energy estimate (2.23) applied to $\tilde{y} - y^\alpha$, we know that

$$|\tilde{y}(T_0) - y^\alpha(T_0)|^2_{L^2(\Omega)} + \int_0^{T_0}\int_\Omega \tilde{\xi}_1^3 e^{-2s\tilde{\eta}_1}|\tilde{y} - y^\alpha|^2 dxdt \leq C\int_0^{T_0}\int_\omega |\tilde{y} - y^\alpha|^2 dxdt.$$

Therefore we obtain, with a different constant C depending on g, w, and q_0,

$$\int_0^{T_0}\int_\omega |\tilde{y} - y^\alpha|^2 dxdt + \alpha|\tilde{y}_0 - y_\alpha|^2_{L^2(\Omega)} \leq C\alpha\left(\int_0^{T_0}\int_\omega |\tilde{y} - y^\alpha|^2 dxdt\right)^{\frac{1}{2}},$$

and this gives immediately the result of Theorem 3.5. □

Of course we obtain completely similar results for the case of linearized Navier–Stokes equations without any additional difficulty.

Acknowledgment. The author wishes to thank the referees for their useful comments and suggestions.

REFERENCES

[1] J. Blum and F. X. Le Dimet, *Assimilation de données pour les fluides géophysiques*, Matapli, 67 (2002), pp. 33–55.

[2] P. Courtier and O. Talagrand, *Variational assimilation of meteorological observations with the adjoint vorticity equation.* I *Theory,* Q. J. R. Meteorol. Soc., 113 (1987), pp. 1311–1328.

[3] A. Dubova, A. Osses, and J.-P. Puel, *Exact controllability on trajectories for a transmission parabolic problem,* ESAIM Control Optim. Calc. Var., 8 (2002), pp. 621–661. Special issue to the memory of Jacques-Louis Lions.

[4] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems,* Kluwer Academic Publishers, Dordrecht, 1996.

[5] H. W. Engl, K. Kunisch, and A. Neubauer, *Convergence rates for Tikhonov regularisation of nonlinear ill-posed problems,* Inverse Problems, 5 (1989), pp. 523–540.

[6] C. Fabre and G. Lebeau, *Prolongement unique des solutions de l'équation de Stokes,* Comm. Partial Differential Equations, 21 (1996), pp. 573–596.

[7] E. Fernandez-Cara, S. Guerrero, O. Yu. Imanuvilov, and J.-P. Puel, *Local exact controllability of the Navier–Stokes system,* J. Math. Pures Appl., 83 (2004), pp. 1501–1542.

[8] A. Fursikov and O. Imanuvilov, *Controllability of Evolution Equations,* Lecture Notes Ser. 34, RIM-GARC, Seoul National University, Seoul, 1996.

[9] G. C. Garcia, A. Osses, and J.-P. Puel, *A data assimilation problem applied to a large-scale ocean circulation model,* submitted.

[10] C. W. Groetsch, *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind,* Res. Notes Math. 105, Pitman, Boston, 1984.

[11] P. Hepperger, *Ein POD-basierter Ansatz zur numerischen Lösung eines Datenassimilationsproblems am Beispiel einer linearen Diffusions-Konvektions-Gleichung,* Diplomarbeit, T.U. München, München, Deutschland, 2007.

[12] O. Imanuvilov, *On exact controllability for the Navier–Stokes equations,* ESAIM Control Optim. Calc. Var., 3 (1998), pp. 97–131.

[13] O. Imanuvilov, *Remarks on exact controllability for Navier–Stokes equations,* ESAIM Control Optim. Calc. Var., 6 (2001), pp. 39–72.

[14] O. Imanuvilov and M. Yamamoto, *Carleman inequalities for parabolic equations in Sobolev spaces of negative order and exact controllability for semilinear parabolic equations,* Publ. Res. Inst. Math. Sci., 39 (2003), pp. 227–274.

[15] F. X. Le Dimet, *Une étude générale d'analyse objective variationnelle des champs météorologiques,* Rapport scientifique LAMP 28, Université de Clermont II, Aubière, France, 1980.

[16] F. X. Le Dimet and O. Talagrand, *Variational algorithms for analysis and assimilation of meteorological observations theoretical aspects,* Tellus Ser. A, 38 (1986), pp. 97–110.

[17] J.-L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations,* Springer-Verlag, Berlin, 1971.

[18] J.-L. Lions, *Quelques méthodes de résolution des problèmes aux limites non linéaires,* Dunod, Gauthier-Villars, Paris, 1969.

[19] S. Mizohata, *Unicité du prolongement des solutions pour quelques opérateurs différentiels paraboliques,* Mem. Coll. Sci. Univ. Kyoto Ser. A, 31 (1958), pp. 219–239.

[20] J.-P. Puel, *Application of Global Carleman Estimates to Controllability and Inverse Problems,* Universidad Federal de Rio de Janeiro, Rio de Janeiro, Brazil, to appear.

[21] J.-P. Puel, *Une approche non classique d'un problème d'assimilation de données,* C. R. Math. Acad. Sci. Paris, 335 (2002), pp. 161–166.

[22] R. Temam, *Navier–Stokes Equations. Theory and Numerical Analysis.* Stud. Math. Appl. 2, North–Holland, Amsterdam, 1977.

SIAM J. Control Optim.
Vol. 48, No. 2, pp. 1112–1133

© 2009 Society for Industrial and Applied Mathematics

A STUDY OF THE ASYMPTOTIC HOLONOMIC EFFICIENCY PROBLEM*

JIANGHAI HU† AND SLOBODAN SIMIC‡

Abstract. In this paper we study an asymptotic version of the holonomic efficiency problem which originated in the study of swimming microorganisms. Given a horizontal distribution on a vector bundle, the holonomy of a loop in the base space is the displacement along the fiber direction of the end points of its horizontal lift. The holonomic efficiency problem is to find the most efficient loop in the base space in terms of gaining holonomy, where the cost of the base loop is measured by a subriemannian metric, and the holonomy gained is compared using a test function. We introduce the notions of rank and asymptotic holonomy and characterize them through the series expansions of holonomy as a function of the loop scale. In the rank two case we prove that for convex test functions the most efficient base loops are simple circles, and we solve these loops for linear and norm test functions. In the higher rank case the analytical solutions are outlined for some special instances of the problem. An example of a turning linked-mass system is worked out in detail to illustrate the results.

Key words. nonholonomic systems, optimal control, holonomy, subriemannian geometry

AMS subject classifications. 53C17, 49K15, 93B18, 70F25, 70G45

DOI. 10.1137/040613895

1. Introduction.

The isoholonomic problem has applications in a variety of fields, for example, the falling cat problem [12] in mechanics, the swimming microorganism at low Reynolds number problem [21] in biology, and the Berry phase problem [17] in quantum mechanics. Generally speaking, for a loop c in the base space M of a principal bundle $\pi : Q \to M$ with a horizontal distribution $\mathcal{H}$, its holonomy is the vertical displacement of the end points of its horizontal lift in Q. The holonomic efficiency of c can then be defined as the ratio of a certain functional of the holonomy and some quantity such as the length or energy that characterizes the cost for traversing c. The isoholonomic problem tries to find the loop c with the highest holonomic efficiency.

In the context of microswimming, various notions of holonomic efficiency have been proposed. To name a few, we mention that in [14] the efficiency of a swimming stroke is defined as the ratio of the square of the average speed achieved by the stroke to the average power output required (this notion of efficiency is scaled by a characteristic thrust to obtain the dimensionless Froude's efficiency studied in [7]), while in [4] the efficiency is the ratio of the product of average speed and a characteristic thrust to the average power. Another notion of efficiency is proposed in [22] that is invariant to temporal and spatial rescaling. See [13] for more discussions on the various notions of efficiency in the microswimming problem, and see [17, 18, 5] for applications in other areas. The isoholonomic problem can be formulated as a special class of optimal control problems, whose solution has been studied, for example, in [3, 6, 20, 8].

*Received by the editors October 22, 2004; accepted for publication (in revised form) October 23, 2008; published electronically March 11, 2009. A part of the results in this paper has appeared in [11].

http://www.siam.org/journals/sicon/48-2/61389.html

†School of Electrical and Computer Engineering, Purdue University, 465 Northwestern Ave., West Lafayette, IN 47907 (jianghai@purdue.edu).

‡Department of Mathematics, San Jose State University, San Jose, CA 95192-0103 (simic@math.sjsu.edu).

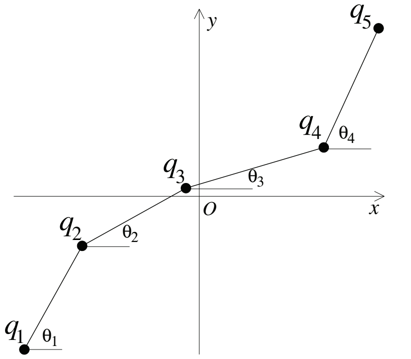


FIG. 1.1. *A linked-mass system with five nodes and four segments.*

There are several limitations in these previous works that motivate the research in this paper. First of all, although all of them deal with the asymptotic case, a rigorous formulation of the asymptotic holonomic efficiency problem has not yet been adequately addressed. Second, the problems studied so far focus on the nondegenerate (rank two) case only, while the degenerate (higher rank) case has been largely ignored. Third, the various notions of efficiency proposed in the literature, with the exception of [17], are defined through linear functionals (test functions) of the holonomy, whereas in some cases general test functions could be more desirable.

In this paper we propose a general framework for studying the asymptotic holonomic efficiency problem. The concepts of rank and asymptotic holonomy are defined, and, using the notion of test functions, we propose the following two optimization problems, which are dual to each other, for studying the most efficient way to gain asymptotic holonomy: the generalized isoholonomic and isoperimetric problems. Both the rank two case and the higher rank case are considered. In the rank two case, we extend the well-known result that optimal solutions are circles from the linear test function case to the general convex test function case, and we propose procedures for finding these circles for norm test functions. For the higher rank case, we focus on a special family of distributions with arbitrary rank and find the optimal solution through optimal control theory.

For simplicity, our problems are formulated on the trivial vector bundle $\pi : \mathbb{R}^{n+k} \to \mathbb{R}^n$. However, due to their asymptotic nature, our results can be easily extended to the case of more general state spaces such as principal bundles.

1.1. Motivating example.

We start by introducing a motivating example, first reported in [11], of a snake-like linked-mass system moving on a plane. A relevant but more complicated model is the molecule model studied in [10].

The system consists of $n+2$ unit point masses (nodes) subsequently connected by $n+1$ rigid links (segments) of unit length and zero mass. Figure 1.1 shows an example with $n = 3$. Given a fixed coordinate system with the origin O, denote by $q_1(t), \ldots, q_{n+2}(t) \in \mathbb{R}^2$ the locations of the $n+2$ nodes at time $t \geq 0$. Assume without loss of generality that $\sum_{i=1}^{n+2} q_i(0) = 0$; i.e., the system is initially centered at the origin O. Suppose that the system is not subject to external forces. Then its total linear momentum and total angular momentum (using the origin O as the center of rotation) are conserved:

$$\sum_{i=1}^{n+2} \dot{q}_i \equiv 0, \tag{1.1}$$

$$\sum_{i=1}^{n+2} q_i \times \dot{q}_i \equiv 0. \tag{1.2}$$

Note that the zero-mass links do not contribute to the above computation.

Condition (1.1) then implies that $\sum_{i=1}^{n+2} q_i \equiv 0$. So the configuration of the system is uniquely determined by the angles $\theta_1, \dots, \theta_{n+1}$, where θ_i is the angle that $q_{i+1} - q_i$ makes with the positive x-axis, $i = 1, \dots, n+1$. Each θ_i takes values in $\mathbb{R}$ modulo 2π, namely, the 1-torus $\mathbb{T} = \mathbb{R}/2\pi\mathbb{Z}$, so $(\theta_1, \dots, \theta_{n+1})$ takes values in the $(n+1)$-torus $\mathbb{T}^{n+1}$, which is the *configuration space* of the system.

Remark 1. There is a natural bundle structure on $\mathbb{T}^{n+1}$. $\mathbb{T}$ acts on $\mathbb{T}^{n+1}$ by

$$R_\theta(\theta_1, \dots, \theta_{n+1}) = (\theta_1 + \theta, \dots, \theta_{n+1} + \theta), \quad \theta \in \mathbb{T},\ (\theta_1, \dots, \theta_{n+1}) \in \mathbb{T}^{n+1}. \tag{1.3}$$

The effect of R_θ on any configuration is a counterclockwise rotation of θ. Each orbit of this action consists of configurations with the same shape but different orientations, and configurations in different orbits have different shapes. Thus the *shape space* of the linked-mass system can be defined as the set of all R-orbits in $\mathbb{T}^{n+1}$, i.e., $\mathbb{T}^{n+1}/\mathbb{T}$, which topologically is an n-torus. The quotient map $\pi : \mathbb{T}^{n+1} \to \mathbb{T}^{n+1}/\mathbb{T}$ defines $\mathbb{T}^{n+1}$ as a $\mathbb{T}$-bundle over $\mathbb{T}^{n+1}/\mathbb{T}$, whose fibers are exactly the R-orbits.

For given $\theta_1, \dots, \theta_{n+1}$, the corresponding $q_1, \dots, q_{n+2}$ satisfying $\sum_{i=1}^{n+2} q_i = 0$ are

$$q_1 = -\frac{1}{n+2} \sum_{j=1}^{n+1} (n+2-j)(\cos\theta_j, \sin\theta_j)^T, \tag{1.4}$$

$$q_i = q_1 + \sum_{j=1}^{i-1} (\cos\theta_j, \sin\theta_j)^T, \quad i = 2, \dots, n+2. \tag{1.5}$$

Equations (1.4) and (1.5) together define an embedding of the configuration space $\mathbb{T}^{n+1}$ into $\mathbb{R}^{2n+4}$. Thus $\mathbb{T}^{n+1}$ inherits isometrically via this embedding a riemannian metric $\langle \cdot, \cdot \rangle$ from the standard metric on $\mathbb{R}^{2n+4}$. From the computation in the appendix, $\langle \cdot, \cdot \rangle$ can be determined as

$$g_{ij} \triangleq \left\langle \frac{\partial}{\partial\theta_i}, \frac{\partial}{\partial\theta_j} \right\rangle = \Delta_{ij} \cos(\theta_i - \theta_j), \quad 1 \le i, j \le n+1, \tag{1.6}$$

where Δ_{ij} are constants defined by

$$\Delta_{ij} = \begin{cases} \frac{i(n+2-j)}{n+2} & \text{if } i < j, \\ \frac{(n+2-i)j}{n+2} & \text{if } i \ge j. \end{cases}$$

Suppose that the trajectory of the linked-mass system over a time interval $I = [0, 1]$ is given by a curve γ in $\mathbb{T}^{n+1}$. Unless otherwise stated, we assume that all curves in this paper are defined on I. Define

$$L(\gamma) = \int_0^1 \|\dot{\gamma}\| \, dt, \qquad E(\gamma) = \int_0^1 \|\dot{\gamma}\|^2 \, dt$$

as the *length* and the *energy* of γ, respectively, where $\|\cdot\|$ is the norm corresponding to $\langle\cdot,\cdot\rangle$. From the definition of $\langle\cdot,\cdot\rangle$, we have $L(\gamma) = \int_0^1 (\sum_{i=1}^{n+2} \|\dot{q}_i\|^2)^{1/2}\, dt$ and $E(\gamma) = \int_0^1 \sum_{i=1}^{n+2} \|\dot{q}_i\|^2\, dt$, where $q_1, \dots, q_{n+2}$ are the positions of the nodes corresponding to γ. A physical explanation of the expression of $E(\gamma)$ is that, since the links in the system have zero mass, their rotations and translations do not contribute to the total energy; hence the energy of the path (maneuver) γ is the time integral of (twice) the total instantaneous kinetic energy of the nodes only, which is derived in the appendix in the alternative coordinates $(\theta_1, \dots, \theta_{n+1})$.

With this metric on $\mathbb{T}^{n+1}$, we now study the geometric implication of the constraint (1.2). It can be verified that a curve $\gamma = (\theta_1, \dots, \theta_{n+1})$ in $\mathbb{T}^{n+1}$ satisfies the constraint (1.2) if and only if

$$\sum_{i,j=1}^{n+1} \Delta_{ij} \cos(\theta_i - \theta_j)\dot{\theta}_j = 0, \tag{1.7}$$

or equivalently, if and only if $\Theta(\dot{\gamma}) = 0$, where Θ is a one-form on $\mathbb{T}^{n+1}$ defined by

$$\Theta = \sum_{i,j=1}^{n+1} \Delta_{ij} \cos(\theta_i - \theta_j) d\theta_j. \tag{1.8}$$

In other words, γ must be a *horizontal* curve for the codimension one distribution $\mathcal{H} \triangleq \ker \Theta$ on $\mathbb{T}^{n+1}$. The restriction of $\langle\cdot,\cdot\rangle$ to $\mathcal{H}$ defines a subriemannian metric $\langle\cdot,\cdot\rangle_{\mathcal{H}}$. In this metric the subriemannian length of a horizontal curve γ is the same as its riemannian length $L(\gamma)$.

We observe that the horizontal distribution $\mathcal{H}$ defined above from (1.7) is naturally induced by the metric $\langle\cdot,\cdot\rangle$ given in (1.6). In fact, at each point $\theta = (\theta_1, \dots, \theta_{n+1}) \in \mathbb{T}^{n+1}$, the codimension one horizontal space $\mathcal{H}_\theta = \ker \Theta_\theta$ can be easily verified as the orthogonal complement of the bundle direction of $\pi : \mathbb{T}^{n+1} \to \mathbb{T}^n$, namely, $(1, \dots, 1)$, under the metric $\langle\cdot,\cdot\rangle$:

$$\mathcal{H}_\theta = \{v \in T_\theta \mathbb{T}^{n+1} \simeq \mathbb{R}^{n+1} \mid \langle v, (1, \dots, 1)\rangle = 0\}.$$

This fact has been observed in the general setting of rotation and vibration motions of molecules in [10]. The connections resulting from such horizontal distributions are called the natural mechanical connections [16].

1.2. Objective and overview of the paper. We are interested in finding the most efficient way for the linked-mass system to turn. More precisely, among all the maneuvers γ that guide the system from a starting configuration $(\theta_1^0, \dots, \theta_{n+1}^0)$ at time 0 to a desired configuration $(\theta_1^0 + \theta, \dots, \theta_{n+1}^0 + \theta)$ at time 1 that has the same shape but a different orientation, subject to the constraint (1.2) of zero total angular momentum, we want to find the one (or ones) with minimal energy (or minimal length $L(\gamma)$, which are equivalent up to reparameterizations). In light of the above discussion, the solutions to this problem are the shortest horizontal curves in $\mathbb{T}^{n+1}$ connecting $(\theta_1^0, \dots, \theta_{n+1}^0)$ to $(\theta_1^0 + \theta, \dots, \theta_{n+1}^0 + \theta)$, which are necessarily distance-minimizing subriemannian geodesics in $\langle\cdot,\cdot\rangle_{\mathcal{H}}$.

Remark 2. In definitions (1.6) and (1.8) the terms involving θ_i's are of the form $\theta_i - \theta_j$, which remain unchanged under the action R. Thus both $\mathcal{H}$ and $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ are invariant along the fibers of $\pi : \mathbb{T}^{n+1} \to \mathbb{T}^{n+1}/\mathbb{T}$, and together they specify a subriemannian geometry invariant with respect to π (see section 2). In this perspective, the problem under study is to determine the shortest horizontal curve connecting two points $(\theta_1, \dots, \theta_{n+1})$ and $R_\theta(\theta_1, \dots, \theta_{n+1})$ in the same fiber.

Unfortunately, solutions to the above formulated problem are usually impossible to obtain analytically due to its global nature (the starting and ending configurations could be far away from each other). In this paper, we shall instead study an asymptotic (local) version of the problem; i.e., what is the most efficient way for the linked-mass system to turn if it can only exert an increasingly small amount of energy? The exact formulation of the asymptotic problem will be given in section 2 in the more general context of codimension k distribution on $\mathbb{R}^{n+k}$. In particular, we define the notions of rank and asymptotic holonomy, and, using test functions, propose an optimization problem that generalizes the efficiency problems studied in the literature. In section 3, we focus on the rank two case and prove that, for convex test functions, at least one of the solutions to the optimization problem is given by a simple circle contained in a two-dimensional plane. Although such solutions are well known in the literature when the test function is linear, our results hold for arbitrary convex test functions. Detailed procedures are also outlined for finding the solution when the test function is a norm. The higher rank case, on the other hand, is much more complicated. In section 4, we solve the problem for a special family of distributions with rank higher than two, and in section 5 we use the result to obtain the asymptotically most efficient maneuver for the linked-mass system in section 1. Section 6 extends the results to principal bundles.

2. Problem formulation.

We now formulate the problem in the general setting of codimension k distributions on the Euclidean space $\mathbb{R}^{n+k}$ for some $n, k \geq 1$. The projection $\pi : (x_1, \dots, x_{n+k}) \in \mathbb{R}^{n+k} \mapsto (x_1, \dots, x_n) \in \mathbb{R}^n$ defines $\mathbb{R}^{n+k}$ as a trivial vector bundle over $\mathbb{R}^n$, whose fiber over each $m \in \mathbb{R}^n$ is given by $\pi^{-1}(m) \simeq \mathbb{R}^k$. We shall first review some relevant concepts in subriemannian geometry. A comprehensive introduction on this topic can be found in [18].

2.1. Codimension k distributions and subriemannian metrics on $\mathbb{R}^{n+k}$.

Let $\Theta = (\Theta_1, \dots, \Theta_k)$ be an $\mathbb{R}^k$-valued one-form on $\mathbb{R}^{n+k}$ with components

$$\Theta_j = dx_{n+j} - \sum_{i=1}^{n} \alpha_i^j dx_i, \quad 1 \leq j \leq k, \tag{2.1}$$

for some C^∞ functions $\alpha_i^j : \mathbb{R}^{n+k} \to \mathbb{R}$. Then $\mathcal{H} = \ker \Theta$ is a codimension k distribution on $\mathbb{R}^{n+k}$. The *horizontal space* $\mathcal{H}_q$ at each $q \in \mathbb{R}^{n+k}$ is the kernel of Θ_q in $T_q\mathbb{R}^{n+k}$, which can be thought of as an n-dimensional subspace of $\mathbb{R}^{n+k}$, i.e., $\mathcal{H}_q = \{(v_1, \dots, v_{n+k}) \in \mathbb{R}^{n+k} : v_{n+j} - \sum_{i=1}^n \alpha_i^j(q) v_i = 0,\ 1 \leq j \leq k\}$. Θ is called the *connection form* of $\mathcal{H}$.

A horizontal curve γ in $\mathbb{R}^{n+k}$ is an absolute continuous curve in $\mathbb{R}^{n+k}$, whose tangent vector $\dot\gamma(t)$ belongs to $\mathcal{H}_{\gamma(t)}$ wherever it exists. Write $\gamma = (\gamma_1, \dots, \gamma_{n+k})$ in coordinates. Then γ is horizontal if and only if $\dot\gamma_{n+j} = \sum_{i=1}^n \alpha_i^j \dot\gamma_i$, $1 \leq j \leq k$, a.e. Fix a pair (m, q), where $m \in \mathbb{R}^n$, $q \in \mathbb{R}^{n+k}$, and $\pi(q) = m$. The *horizontal lift* (based at q) of a curve c in $\mathbb{R}^n$ starting from m is defined as the unique horizontal curve γ in $\mathbb{R}^{n+k}$ starting from q and satisfying $\pi(\gamma) = c$. If $c = (c_1, \dots, c_n)$ in coordinates, then $\gamma = (\gamma_1, \dots, \gamma_{n+k})$ is obtained by solving the following differential equations:

$$\gamma_1 = c_1, \dots, \gamma_n = c_n,\ \dot\gamma_{n+j} = \sum_{i=1}^{n} \alpha_i^j(\gamma) \dot c_i, \qquad 1 \leq j \leq k. \tag{2.2}$$

If, in particular, c is a close loop, then γ starts and ends in the same fiber $\pi^{-1}(m)$; i.e., $\gamma(1)-\gamma(0)$ is of the form $(0,\dots,0,h)$ for some $h \in \mathbb{R}^k$. We called h the *holonomy* of the loop c, which in general depends on the base point $q \in \pi^{-1}(m)$ of γ.

A subriemannian metric $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ on $\mathcal{H}$ is a smooth assignment of inner products to the horizontal spaces $\mathcal{H}_q$. The length of a horizontal curve γ is measured as $L(\gamma) = \int_0^1 \|\dot\gamma\|_{\mathcal{H}}\, dt = \int_0^1 \langle\dot\gamma,\dot\gamma\rangle_{\mathcal{H}}^{1/2}\, dt$ under this metric. The subriemannian distance between two points in $\mathbb{R}^{n+k}$ is the infimum of the length of all horizontal curves connecting them. Thus $\mathcal{H}$ and $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ specify a subriemannian geometry on $\mathbb{R}^{n+k}$.

2.2. Invariant distributions and subriemannian metrics. The distribution $\mathcal{H}$ is called π-*invariant*, or simply invariant, with the bundle structure $\pi : \mathbb{R}^{n+k} \to \mathbb{R}^n$ if its horizontal spaces are invariant along fibers. In terms of (2.1), this is equivalent to

$$\alpha_i^j(x_1,\dots,x_{n+k}) = \alpha_i^j(x_1,\dots,x_n), \quad 1 \le i \le n,\ 1 \le j \le k. \tag{2.3}$$

So we can think of α_i^j as functions on $\mathbb{R}^n$ and define an $\mathbb{R}^k$-valued one-form on $\mathbb{R}^n$ as

$$\alpha = \begin{bmatrix}\alpha_1\\ \vdots \\ \alpha_k\end{bmatrix} \triangleq \begin{bmatrix}\sum_{i=1}^n \alpha_i^1 dx_i\\ \vdots \\ \sum_{i=1}^n \alpha_i^k dx_i\end{bmatrix}. \tag{2.4}$$

The holonomy of a loop c in $\mathbb{R}^n$ is then independent of the starting point of its horizontal lift γ, and thus can be simply denoted by $h(c)$. Indeed, by (2.2) and an application of the Stokes' theorem,

$$h(c) = \begin{bmatrix}\int_0^1 \dot\gamma_{n+1}\, dt\\ \vdots \\ \int_0^1 \dot\gamma_{n+k}\, dt\end{bmatrix} = \int_c \alpha = \int_S d\alpha = \int_S \beta, \tag{2.5}$$

where S is a two-dimensional submanifold immersed in $\mathbb{R}^n$ whose boundary ∂S is exactly c, and β is the $\mathbb{R}^k$-valued two-form defined by

$$\beta \triangleq d\alpha = \sum_{1\le i,j\le n} \beta_{ij}\, dx_i \wedge dx_j, \tag{2.6}$$

where β_{ij}, $1 \le i,j \le n$, are $\mathbb{R}^k$-valued functions on $\mathbb{R}^n$ with $\beta_{ij} = -\beta_{ji}$. In (2.5), $h(c)$ is written as an integral of β over an arbitrary surface encircled by c.

For an invariant distribution $\mathcal{H}$, a subriemannian metric $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ is called π-*invariant* (or invariant) with the bundle structure $\pi : \mathbb{R}^{n+1} \to \mathbb{R}^n$ if it is also invariant along fibers. Invariant subriemannian metrics $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ on $\mathcal{H}$ have a one-to-one correspondence with riemannian metrics $\langle\cdot,\cdot\rangle_{\mathbb{R}^n}$ on the base space $\mathbb{R}^n$ according to the following relation:

$$\langle h_q(u), h_q(v)\rangle_{\mathcal{H}} = \langle u,v\rangle_{\mathbb{R}^n}, \quad \forall u,v \in T_m\mathbb{R}^n,\ m \in \mathbb{R}^n,\ q \in \pi^{-1}(m). \tag{2.7}$$

Here $h_q : T_m\mathbb{R}^n \to \mathcal{H}_q$ is the horizontal lift operator defined as the inverse map of the linear isomorphism $d\pi_q : \mathcal{H}_q \to \mathbb{T}_m\mathbb{R}^n$. We call $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ satisfying (2.7) the horizontal lift of $\langle\cdot,\cdot\rangle_{\mathbb{R}^n}$.

2.3. Asymptotic holonomy. Let $\mathcal{H} = \ker\Theta$ be a codimension k distribution on $\mathbb{R}^{n+k}$ with the connection form Θ given in (2.1), and let $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ be a subriemannian metric on $\mathcal{H}$. In the rest of this paper, we shall assume that both $\mathcal{H}$ and $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ are invariant with the bundle structure $\pi : \mathbb{R}^{n+k} \to \mathbb{R}^n$. Thus we can define the forms α and β as in (2.4) and (2.6), and $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ is the horizontal lift of a metric $\langle\cdot,\cdot\rangle_{\mathbb{R}^n}$ on the base space $\mathbb{R}^n$. It should be pointed out, however, that the concepts of asymptotic rank and efficiency and some of their properties described below can be easily generalized to the noninvariant case.

Fix a point $m \in \mathbb{R}^n$ and a loop $c \not\equiv 0$ in $\mathbb{R}^n$ based at m. For each $\epsilon > 0$, denote by $c_\epsilon = m + \epsilon(c - m)$ the loop based at m obtained by scaling c by a factor of ϵ towards m, and let γ_ϵ be the horizontal lift of c_ϵ in $\mathbb{R}^{n+k}$ based at $q \in \pi^{-1}(m) \subset \mathbb{R}^{n+k}$. We can define the following two quantities for c_ϵ: (1) its length $L(c_\epsilon) > 0$ is the length of c_ϵ in $\mathbb{R}^n$ as measured by the metric $\langle\cdot,\cdot\rangle_{\mathbb{R}^n}$, which by (2.7) is also the length of the horizontal curve γ_ϵ as measured by $\langle\cdot,\cdot\rangle_{\mathcal{H}}$; (2) its holonomy $h(c_\epsilon) \in \mathbb{R}^k$ is the vertical displacement between the two end points of γ_ϵ.

It is easy to see that $L(c_\epsilon)$ is of the same order of ϵ as $\epsilon \to 0$.

LEMMA 1. *$L(c_\epsilon) = a\epsilon + o(\epsilon)$, where $a \neq 0$ depends on $\langle\cdot,\cdot\rangle_{\mathbb{R}^n}$ only through its restriction at m.*

The continuity of $\langle\cdot,\cdot\rangle_{\mathbb{R}^n}$ is needed to show the above claim. As for $h(c_\epsilon)$, we have the following.

LEMMA 2. *$h(c_\epsilon) = \epsilon^{r(m)}\hat{h}(c) + o(\epsilon^{r(m)})$ for some constant $\hat{h}(c) \in \mathbb{R}^k$ and an integer*

$$r(m) = \min\{l : \textit{at least one } l\textit{th order partial derivative of } \beta \textit{ at } m \textit{ is nonzero}\} + 2. \tag{2.8}$$

Moreover, $\hat{h}(c) \neq 0$ for at least one loop c based at m.

Remark 3. The lth order partial derivatives of β at m are terms of the form $\frac{\partial^l \beta(m)}{\partial x_1^{l_1}\cdots\partial x_n^{l_n}}$ for some integers $l_1, \dots, l_n$ with $l_1 + \cdots + l_n = l$. Taking values in the set of $\mathbb{R}^k$-valued skew-symmetric 2-tensors on $\mathbb{R}^n$, each of these terms is zero if and only if all of its k components are zero.

Proof. Let S be a two-dimensional submanifold of $\mathbb{R}^n$ encircled by c. For each $\epsilon > 0$, denote $S_\epsilon = m + \epsilon(S - m)$. Then $\partial S_\epsilon = c_\epsilon$ and $S_\epsilon \to m$ as $\epsilon \to 0$. Expanding $\beta = \sum_{1\le i,j\le n} \beta_{ij} dx_i \wedge dx_j$ at $m = (m_1, \dots, m_n)$ in Taylor expansions and noticing the definition of $r(m)$ in (2.8), we have, for $x \in S_\epsilon$,

$$\beta(x) = \sum_{l_1+\cdots+l_n = r(m)-2} \sum_{1\le i,j\le n} \frac{\partial^{r(m)-2}\beta_{ij}(m)}{\partial x_1^{l_1}\cdots\partial x_n^{l_n}} \frac{(x_1-m_1)^{l_1}\cdots(x_n-m_n)^{l_n}}{l_1!\cdots l_n!} dx_i \wedge dx_j + o(\epsilon^{r(m)-2}).$$

So the dominating term of $h(c_\epsilon) = \int_{S_\epsilon} \beta$ as $\epsilon \to 0$ is

$$\sum_{l_1+\cdots+l_n = r(m)-2} \sum_{1\le i,j\le n} \frac{1}{l_1!\cdots l_n!} \frac{\partial^{r(m)-2}\beta_{ij}(m)}{\partial x_1^{l_1}\cdots\partial x_n^{l_n}} \int_{S_\epsilon} (x_1-m_1)^{l_1}\cdots(x_n-m_n)^{l_n}\, dx_i \wedge dx_j,$$

which is exactly of the form $\epsilon^{r(m)}\hat{h}(c)$, where $\hat{h}(c)$ is given by

$$\sum_{l_1+\cdots+l_n = r(m)-2} \sum_{1\le i,j\le n} \frac{1}{l_1!\cdots l_n!} \frac{\partial^{r(m)-2}\beta_{ij}(m)}{\partial x_1^{l_1}\cdots\partial x_n^{l_n}} \int_{S} (x_1-m_1)^{l_1}\cdots(x_n-m_n)^{l_n}\, dx_i \wedge dx_j. \tag{2.9}$$

It is easy to see that $\hat{h}(c) \neq 0$ for suitably chosen c and S. □

Definition 1. *$r(m)$ defined in* (2.8) *is called the* rank *of $\mathcal{H}$ at $m \in M$, and*

$$\eta(c) \triangleq \frac{\hat{h}(c)}{L^{r(m)}(c)} \in \mathbb{R}^k \tag{2.10}$$

is called the asymptotic holonomy *of the loop c based at m, with $\hat{h}(c)$ defined in* (2.9).

Despite its definition process, the rank $r(m)$ does not depend on the loop c. Indeed, by (2.8), $r(m)$ does not even depend on the subriemannian metric $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and is solely determined by the distribution $\mathcal{H}$ on the fibers over a neighborhood of m. Thus $r(m)$ is an intrinsic quantity of $\mathcal{H}$. On the other hand, the asymptotic holonomy $\eta(c)$ does depend on c, and by Lemma 1 it is also affected by $\langle\cdot,\cdot\rangle_{\mathbb{R}^n}$ (hence by $\langle\cdot,\cdot\rangle_{\mathcal{H}}$) through its restriction at m. Let $A \in \mathbb{R}^{n\times n}$ be the positive definite matrix corresponding to the restriction of $\langle\cdot,\cdot\rangle_{\mathbb{R}^n}$ at m, i.e.,

$$\langle u, v\rangle_{\mathbb{R}^n} = u^T A v \tag{2.11}$$

for all $u, v \in T_m\mathbb{R}^n$. Then to compute $\eta(c)$ one can assume for convenience and without loss of generality that $\langle\cdot,\cdot\rangle_{\mathbb{R}^n}$ is given by A uniformly on $\mathbb{R}^n$; i.e., (2.11) holds for $u, v \in T_x\mathbb{R}^n$ for arbitrary $x \in \mathbb{R}^n$. Finally, the distribution $\mathcal{H}$ affects $\eta(c)$ through $\hat{h}(c)$. By (2.9), in terms of computing $\eta(c)$, the form $\beta = \sum_{1\le i,j\le n} \beta_{ij}\, dx_i \wedge dx_j$ can be replaced by the first nonvanishing term of its Taylor expansions: $\sum_{1\le i,j\le n}$ $\sum_{l_1+\cdots+l_n=r(m)-2} \frac{1}{l_1!\cdots l_n!} \frac{\partial^{r(m)-2}\beta_{ij}(m)}{\partial x_1^{l_1}\cdots\partial x_n^{l_n}}\, dx_i \wedge dx_j$; i.e., we can assume that the components of $\beta_{ij}(x)$ are homogeneous polynomials of degree $r(m)-2$ in x with constant coefficients.

A direct consequence of Lemma 2 and Definition 1 is that, for any loop c based at m with $\eta(c) \neq 0$, we have $h(c_\epsilon) \sim \eta(c)[\epsilon L(c)]^{r(m)}$ as $\epsilon \to 0$. Here $a(\epsilon) \sim b(\epsilon)$ means that $\lim_{\epsilon\to 0} a(\epsilon)/b(\epsilon) = 1$ for functions a and b of $\epsilon > 0$ satisfying $b(\epsilon) \neq 0$ for $\epsilon \neq 0$.

Since $\hat{h}(\cdot)$ and $L^{r(m)}(\cdot)$ are both homogeneous of degree $r(m)$ in the scale ϵ of c_ϵ and are both invariant to reparameterizations of c, $\eta(c)$ has the following properties.

Lemma 3 (invariance of asymptotic holonomy). *The asymptotic holonomy $\eta(c)$ of a loop c based at m is invariant to both scalings and reparameterizations of c, i.e.,*

- *$\eta(c) = \eta(c_\epsilon)$ for any $\epsilon > 0$;*
- *$\eta(c \circ \rho) = \eta(c)$ for any orientation-preserving diffeomorphism $\rho : I \to I$.*

As a result, $\eta(c)$ is a function of only the shape of the curve traversed by c, not of its size or the speed at which it is traversed. Indeed, $\eta(c)$ also remains unchanged if c is defined on a time interval $[0, T]$ rather than $[0, 1]$. As argued in [22], these invariance properties are essential for a meaningful definition of the notion of holonomic efficiency. On the other hand, $\eta(c)$ changes sign if c is traversed in the reverse direction.

2.4. Optimization problem. In this section we define an optimization problem, generalizing the one proposed in section 1.

Definition 2. *A test function $\mathcal{F}$ is a continuous map from $\mathbb{R}^k$ to $\mathbb{R}$ such that $\mathcal{F}(0) = 0$ and $\mathcal{F}$ is linear along rays starting from the origin: $\mathcal{F}(\mu x) = \mu\mathcal{F}(x)$ for all $x \in \mathbb{R}^k, \mu \geq 0$.*

The two important classes of $\mathcal{F}$ considered in this paper are (i) linear functions $\mathcal{F}(x) = \lambda^T x$ for some $\lambda \in \mathbb{R}^k$, and (ii) $\mathcal{F}(x) = \|x\|$ for some norm $\|\cdot\|$ on $\mathbb{R}^k$.

Problem 1. *Find the loop c in M based at $m \in M$ maximizing $\mathcal{F}[\eta(c)]$.*

Problem 1 originates as follows. Let $V : \mathbb{R}^k \to \mathbb{R}$ be a function with $V(0) = 0$ such that $V[h(c)]$ can be interpreted as the performance measure of the loop c. Then Problem 1 is the asymptotic version of the problem of finding the best performing loops c. Indeed, define the best performing c in the asymptotic sense as the ones for which $V[h(c_\epsilon)] \sim \mu[L(c_\epsilon)]^r$ as $\epsilon \to 0$ for the largest possible $\mu \in \mathbb{R}$ and the smallest possible integer r. Since $h(c_\epsilon) \sim \eta(c)[L(c_\epsilon)]^{r(m)}$ by the discussion in section 2.3, we have $V[h(c_\epsilon)] \sim \mathcal{F}[\eta(c)][L(c_\epsilon)]^{r(m)}$ as $\epsilon \to 0$, where $\mathcal{F} : \mathbb{R}^k \to \mathbb{R}$ is defined as

$$\mathcal{F}(h) \triangleq \lim_{\epsilon \to 0^+} \frac{1}{\epsilon} V(\epsilon h), \quad h \in \mathbb{R}^k. \tag{2.12}$$

Hence, in the expression $V[h(c_\epsilon)] \sim \mu[L(c_\epsilon)]^r$, the smallest possible r is $r(m)$ and the largest possible μ is $\max\{\mathcal{F}[\eta(c)] : c\}$, both of which are achieved by solutions c to Problem 1, provided that the optimal $\mathcal{F}[\eta(c)] \neq 0$. Since normally V is differentiable along rays emitting from the origin, $\mathcal{F}$ in (2.12) is well defined and satisfies the conditions in Definition 2. In particular, $\mathcal{F}$ is linear if V is differentiable at 0, and $\mathcal{F} = V$ if V is a norm on $\mathbb{R}^k$.

In Problem 1 we assume that $\mathcal{F}[\eta(c)] > 0$ for at least one c to exclude the trivial solution $c \equiv 0$. This assumption is always satisfied for the examples in this paper.

Since $\mathcal{F}[\eta(c)] = \mathcal{F}[\hat{h}(c)]/L^{r(m)}(c)$, and both $\mathcal{F}[\hat{h}(\cdot)]$ and $L^{r(m)}(\cdot)$ are homogeneous of degree $r(m)$ in the scale ϵ of c_ϵ, solving Problem 1 is equivalent to solving any of the following variational problems:

(2.13) Find c that minimize $L(c)$ subject to $\mathcal{F}[\hat{h}(c)] = 1$;

(2.14) find c that maximize $\mathcal{F}[\hat{h}(c)]$ subject to $L(c) = 1$;

(2.15) find c that minimize $E(c)$ subject to $\mathcal{F}[\hat{h}(c)] = 1$;

(2.16) find c that maximize $\mathcal{F}[\hat{h}(c)]$ subject to $E(c) = 1$.

Here $E(c) \triangleq \int_0^1 \|\dot{c}\|^2_{\mathbb{R}^n}\, dt$ is the energy of c. Problems (2.14) and (2.16) are dual to problems (2.13) and (2.15), respectively. That problems (2.13) and (2.15), and hence problems (2.14) and (2.16), are equivalent is because of the inequality $E(c) \geq L^2(c)$ with equality if and only if c has constant speed; thus solutions to the latter are necessarily solutions to the former parameterized with constant speed. In this paper, to avoid the ambiguity of parameterizations, we will study problems (2.15) and (2.16).

By the discussion immediately after Definition 1, to solve these problems we can assume the following without loss of generality.

Assumption 1. *Assume that*

1. $m = 0$ *is the origin of* $\mathbb{R}^n$*;*
2. $\langle \cdot, \cdot \rangle_{\mathbb{R}^n}$ *is given by a positive definite* $A \in \mathbb{R}^{n\times n}$ *on* $\mathbb{R}^n$*. In fact, after a change of orthonormal coordinates, we can assume that* $A = I_n$*; i.e.,* $\langle \cdot, \cdot \rangle_{\mathbb{R}^n}$ *is the standard metric on* $\mathbb{R}^n$*;*
3. $\beta = \sum_{1\leq i,j\leq n} \beta_{ij} dx_i \wedge dx_j \neq 0$*, where* $\beta_{ij} = -\beta_{ji}$*, and the components of* $\beta_{ij}(x) \in \mathbb{R}^k$ *are homogeneous polynomials of degree* $r(0) - 2$ *in* $x \in \mathbb{R}^n$ *with constant coefficients.*

Thus $L(c)$ and $E(c)$ are the standard arc length and energy, respectively, of the loop c based at 0, and $\hat{h}(c)$ in (2.9) coincides with $h(c)$. Problems (2.15) and (2.16) can then be reformulated, respectively, as the following.

Problem 2 (generalized isoholonomic problem). *Find* c *with* $\mathcal{F}[h(c)] = 1$ *minimizing* $E(c)$*.*

Problem 3 (generalized isoperimetric problem). *Find* c *with* $E(c) = 1$ *maximizing* $\mathcal{F}[h(c)]$*.*

For an even more general formulation of the isoholonomic problem, see [17].

Solutions to the above two problems are the same up to a scaling. To derive their equations, note that the solutions to Problem 2 also solve the following problem for some proper $h_0 \in \mathbb{R}^k$:

$$\text{Find } c \text{ with a fixed } h(c) = h_0 \text{ that minimize } E(c). \tag{2.17}$$

It is shown in [18] that the solutions to problem (2.17) satisfy

$$\ddot{c} = -i_{\dot{c}}(\lambda^T \beta) \tag{2.18}$$

for some constant $\lambda \in \mathbb{R}^k$. Here $\lambda^T \beta$ is an $\mathbb{R}$-valued two-form, and $i_{\dot{c}}(\lambda^T \beta) \triangleq \lambda^T \beta(\dot{c}, \cdot)$ is a one-form on $\mathbb{R}^n$ that we identify as a vector in $\mathbb{R}^n$ via the canonical metric. Write $\beta = \sum_{1 \leq i,j \leq n} \beta_{ij}\, dx_i \wedge dx_j$ in coordinates. Then (2.18) is equivalent to

$$\ddot{c} = Z\dot{c}, \tag{2.19}$$

where Z is the skew-symmetric matrix

$$Z = \begin{bmatrix} 2\lambda^T \beta_{11} & \cdots & 2\lambda^T \beta_{1n} \\ \vdots & \vdots & \vdots \\ 2\lambda^T \beta_{n1} & \cdots & 2\lambda^T \beta_{nn} \end{bmatrix}, \tag{2.20}$$

whose components are homogeneous polynomials of degree $r(0) - 2$ in x with constant coefficients. Indeed, (2.19) describes the motions of a particle of unit mass and unit charge moving in a magnetic field given by Z on $\mathbb{R}^n$ when $n = 2, 3$ (see [2, 17]). Equation (2.19) can also be derived from the Pontryagin maximum principle [19] by formulating problem (2.17) as an optimal control problem.

Equation (2.19) does not solve Problem 2 completely, as λ is unknown and we are interested only in those solutions that start and end in the origin. Thus we still need to determine λ and the appropriate initial condition $\dot{c}(0)$ such that $c(0) = c(1) = 0$, which is often a nontrivial task.

3. Rank two case.

We first study the solution of Problem 2 in the simplest case, namely, the rank two case. In this case, when the test function is linear, it is a well-known fact that the optimal loops are simple circles, and straightforward procedures exist to find these optimal circles (see section 3.1). Indeed, when $n = 2$, the rank two isoholonomic problem degenerates into the classical isoperimetric problem. See the discussion in [18, Chap. 1]. Our contribution in this section lies in the generalization of this result to the case of convex test functions (see Theorem 1). In addition, for a class of convex but not linear test functions, in section 3.2 we propose iterative procedures to solve for the optimal circles, which is a much more difficult problem than in the linear test function case.

Suppose that $r(0) = 2$. Then $\beta = \sum_{1 \leq i,j \leq n} \beta_{ij}\, dx_i \wedge dx_j \neq 0$ for constants $\beta_{ij} = -\beta_{ji}$. Z defined in (2.20) is a constant matrix. Being skew-symmetric, Z admits a decomposition of the form

$$Z = Q \cdot \operatorname{diag}\left(\begin{bmatrix} 0 & -\sigma_1 \\ \sigma_1 & 0 \end{bmatrix}, \ldots, \begin{bmatrix} 0 & -\sigma_l \\ \sigma_l & 0 \end{bmatrix}, 0, \ldots, 0 \right) \cdot Q^T,$$

where $Q \in \mathbb{R}^{n \times n}$ is orthonormal, and $\sigma_1 \geq \cdots \geq \sigma_l > 0$ for some integer l with $2l = \operatorname{rank}(Z)$. After an orthonormal coordinate transformation $y = Q^T x$, (2.19) becomes

$$\begin{cases} \begin{bmatrix} \ddot{y}_{2p-1} \\ \ddot{y}_{2p} \end{bmatrix} = \begin{bmatrix} 0 & -\sigma_p \\ \sigma_p & 0 \end{bmatrix} \begin{bmatrix} \dot{y}_{2p-1} \\ \dot{y}_{2p} \end{bmatrix}, & p = 1, \ldots, l, \\ \ddot{y}_p = 0, & p = 2l+1, \ldots, n. \end{cases}$$

Solutions to this equation that start and end at the origin are necessarily of the form

$$\begin{aligned} c(t) = [&a_1(1-\cos(2n_1\pi t)), -a_1\sin(2n_1\pi t), \dots, a_l(1-\cos(2n_l\pi t)),\\ &-a_l\sin(2n_l\pi t), 0, \dots, 0]^T \end{aligned} \tag{3.1}$$

for some $a_1, \dots, a_l \in \mathbb{R}$ and some $n_1, \dots, n_l \in \mathbb{N}$ with $\sigma_1 = 2n_1\pi, \dots, \sigma_l = 2n_l\pi$. Note that we can assume without loss of generality that n_p, $p = 1, \dots, l$, are all distinct. Otherwise, for example, if $n_1 = n_2$, then a suitable change of orthonormal coordinates within the 4-subspace spanned by the $y_1, \dots, y_4$ axes can transform c into the form

$$[a(1-\cos(2n_1\pi t)), -a\sin(2n_1\pi t), 0, 0, a_3(1-\cos(2n_3\pi t)), -a_3\sin(2n_3\pi t), \dots]^T$$

with $a = \sqrt{a_1^2 + a_2^2}$. This step can be repeated until all n_p are eventually distinct, resulting in a curve of the form

$$c \text{ is given in (3.1) for some } 1 \le l \le [n/2] \text{ and distinct } n_1, \dots, n_l \neq 0. \tag{3.2}$$

Curves of the form (3.2) in some orthonormal coordinates of $\mathbb{R}^n$ are called *mixed circles.* If, in particular, $l = 1$ and $n_1 = 1$ in (3.2), the resulting curves are called *simple circles*, which are planar circles in $\mathbb{R}^n$ traversed exactly once.

It is seen from the above that the solutions to Problem 2 are mixed circles. In the case of convex $\mathcal{F}$, the solution can be further simplified.

THEOREM 1. *Suppose that $\mathcal{F} : \mathbb{R}^k \to \mathbb{R}$ is convex. Then there is at least one simple circle solution to Problem* 2 *(respectively, Problem* 3*).*

The result of Theorem 1 is well known in the literature for the case when $\mathcal{F}$ is a linear function. To prove it in the general convex $\mathcal{F}$ case, we first introduce an intermediate result, which is reformulated from the arguments in [18, Sec. 12.3.5]. Consider a mixed circle c of the form (3.2) in some orthonormal coordinates $(y_1, \dots, y_n)$. For each $p = 1, \dots, l$, denote by $c_{(p)}$ the orthogonal projection of c onto the plane spanned by the y_{2p-1} and y_{2p} axes, which is a planar circle traversed n_p times.

LEMMA 4. $h(c) = h(c_{(1)}) + \cdots + h(c_{(l)})$.

Proof. Write $\beta = \sum_{1\le i,j\le n} \hat\beta_{ij}\, dy_i \wedge dy_j$ in the new coordinates, with constants $\hat\beta_{ij} = -\hat\beta_{ji}$. Define $\alpha \triangleq \sum_{i,j=1}^n \hat\beta_{ij} y_i dy_j$. Then $d\alpha = \beta$, and, by (2.5), $h(c) = \int_c \alpha = \sum_{i,j=1}^n \hat\beta_{ij} \int_c y_i\, dy_j$. Note that because of the special form of c in (3.1) and (3.2), unless $\{i,j\} = \{2p-1, 2p\}$ for some $p = 1, \dots, l$, we must have $\int_c y_i\, dy_j = 0$, since the integral of the product of two periodic sine or cosine functions with different frequencies is zero. As a result,

$$h(c) = \sum_{p=1}^{l} \int_c \hat\beta_{2p-1,2p}(y_{2p-1}dy_{2p} - y_{2p}dy_{2p-1}) = \sum_{p=1}^{l} \int_{c_{(p)}} \alpha = \sum_{p=1}^{l} h(c_{(p)}),$$

which proves the desired conclusion. □

Now define three subsets of $\mathbb{R}^k$:

$$\begin{aligned} \mathcal{B}_0 &= \{h(c) : \ c \text{ is a loop with } E(c) \le 1\},\\ \mathcal{B}_1 &= \{h(c) : \ c \text{ is a mixed circle with } E(c) \le 1\},\\ \mathcal{B}_2 &= \{h(c) : \ c \text{ is a simple circle with } E(c) \le 1\}. \end{aligned}$$

$\mathcal{B}_0$ is the set of holonomy achievable by loops with energy no larger than 1 and is the intersection of the unit subriemannian ball centered at 0 with the fiber $\mathbb{R}^k$ through 0.

Obviously, $\mathcal{B}_0$ is star-shaped ($h \in \mathcal{B}_0$ implies $\mu h \in \mathcal{B}_0$ for $\mu \in [0,1]$) and symmetric ($h \in \mathcal{B}_0$ implies $-h \in \mathcal{B}_0$).

Since our previous analysis shows that every holonomy achievable by a loop c can be achieved by a mixed circle with no more energy, we have $\mathcal{B}_0 = \mathcal{B}_1$. Obviously, $\mathcal{B}_2 \subset \mathcal{B}_1$. But Lemma 4 implies the following.

LEMMA 5. *Co*$(\mathcal{B}_1) =$*Co*$(\mathcal{B}_2)$*; i.e.,* $\mathcal{B}_1$ *and* $\mathcal{B}_2$ *span the same convex hull.*

Proof. Since $\mathcal{B}_1$ and $\mathcal{B}_2$ are closed sets, it suffices to show that for any $\lambda \in \mathbb{R}^k$, $\sup(\lambda^T \mathcal{B}_1) = \sup(\lambda^T \mathcal{B}_2)$. Suppose that $\sup(\lambda^T \mathcal{B}_1)$ is achieved at $h(c) \in \mathcal{B}_1$ for a mixed circle c of the form (3.2) in some orthonormal coordinates with $a_1, \ldots, a_l \neq 0$ and $E(c) \leq 1$. Note that $\lambda^T h(c) \geq 0$ since $0 \in \mathcal{B}_1$. By Lemma 4, $h(c) = h(c_{(1)}) + \cdots + h(c_{(l)})$, and

$$\frac{\lambda^T h(c)}{E(c)} = \frac{\lambda^T h(c_{(1)}) + \cdots + \lambda^T h(c_{(l)})}{E(c_{(1)}) + \cdots + E(c_{(l)})} \leq \max_{1\leq p\leq l} \frac{\lambda^T h(c_{(p)})}{E(c_{(p)})}.$$

Suppose that the maximum in the above equation is achieved at p. Then the simple circle $\hat{c}_{(p)}(t) = \frac{1}{2\pi}(0, \ldots, 0, 1-\cos(2\pi t), \sin(2\pi t), 0, \ldots, 0)$ traversing the (scaled) image of $c_{(p)}$ exactly once has unit energy and holonomy $h(\hat{c}_{(p)}) = n_p h(c_{(p)})/E(c_{(p)})$. From the above inequality, we have

$$\lambda^T h(\hat{c}_{(p)}) = n_p \lambda^T h(c_{(p)})/E(c_{(p)}) \geq \lambda^T h(c_{(p)})/E(c_{(p)}) \geq \lambda^T h(c)/E(c) \geq \lambda^T h(c) \geq 0.$$

Since $h(\hat{c}_{(p)}) \in \mathcal{B}_2 \subset \mathcal{B}_1$, $\sup(\lambda^T \mathcal{B}_2) = \sup(\lambda^T \mathcal{B}_1)$. Therefore, Co$(\mathcal{B}_1) =Co(\mathcal{B}_2)$. □

Example 1 (Brockett [6]). Consider the total space $\mathbb{R}^n \oplus \mathfrak{so}_n \simeq \mathbb{R}^{n(n+1)/2}$, whose elements are (x, A) with $x \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$ skew-symmetric matrices. Let $\mathcal{H}$ be the codimension $\frac{n(n-1)}{2}$ distribution invariant with $\pi : \mathbb{R}^n \oplus \mathfrak{so}_n \to \mathbb{R}^n$ given by the $\mathfrak{so}_n$-valued one-form $\alpha = (x \cdot dx^T - dx \cdot x^T)/2$. Thus $\beta = dx \wedge dx^T$. It is easy to verify that $\mathcal{B}_1$ consists of all matrices of the form

$$Q \cdot \mathrm{diag}\left(\begin{bmatrix} 0 & n_1\pi a_1^2 \\ -n_1\pi a_1^2 & 0 \end{bmatrix}, \ldots, \begin{bmatrix} 0 & n_l\pi a_l^2 \\ -n_l\pi a_l^2 & 0 \end{bmatrix}, 0, \ldots, 0 \right) \cdot Q^T$$

for some $Q \in \mathbb{O}_n$, $1 \leq l \leq [\frac{n}{2}]$, $a_1, \ldots, a_l \in \mathbb{R}$, and some distinct $n_1, \ldots, n_l \in \mathbb{N}$ such that $4n_1\pi^2 a_1^2 + \cdots + 4n_l\pi^2 a_l^2 \leq 1$. On the other hand, $\mathcal{B}_2$ is

$$\mathcal{B}_2 = \left\{ Q \cdot \mathrm{diag}\left(\begin{bmatrix} 0 & \pi a_1^2 \\ -\pi a_1^2 & 0 \end{bmatrix}, 0, \ldots, 0 \right) \cdot Q^T : \ Q \in \mathbb{O}_n,\, 4\pi^2 a_1^2 \leq 1 \right\}.$$

Note that $\mathcal{B}_1 \neq \mathcal{B}_2$ since matrices in $\mathcal{B}_2$ have rank at most two while the rank of matrices in $\mathcal{B}_1$ can be any even number between zero and n. In other words, certain holonomy in $\mathfrak{so}_n$ can be achieved by mixed circles but not simple circles.

This example is universal for the problem studied in this section: Any other distribution invariant with $\pi : \mathbb{R}^{n+k} \to \mathbb{R}^n$ specified by a form β with nontrivial constant coefficients is induced from $\mathcal{H}$ in this example by a linear transformation $\mathbb{R}^n \oplus \mathfrak{so}_n \to \mathbb{R}^{n+k}$ that leaves $\mathbb{R}^n$ invariant and properly transforms $\mathfrak{so}_n$ to $\mathbb{R}^k$. Therefore, as an alternative it suffices to prove Lemma 5 for $\mathcal{B}_1$ and $\mathcal{B}_2$ in this particular example only, since convexity is preserved by linear transformations.

Theorem 1 then follows easily from Lemma 5. In fact, Problem 3 is equivalent to finding $\max\{\mathcal{F}(h) : h \in \mathcal{B}_0\} = \max\{\mathcal{F}(h) : h \in \mathcal{B}_1\}$. By Lemma 5 and the convexity of $\mathcal{F}$, $\max\{\mathcal{F}(h) : h \in \mathcal{B}_1\} = \max\{\mathcal{F}(h) : h \in \mathcal{B}_2\}$. So there is at least a simple circle solution to Problem 3. Since solutions to Problem 2 are scaled versions of solutions to Problem 3, this proves Theorem 1.

Remark 4. For the above reasoning to hold, we only need $\mathcal{F}$ to be *quasi convex* instead of convex (a function $f : S \to \mathbb{R}$ defined on a convex subset of $\mathbb{R}^n$ is called quasi convex if each of its sublevel sets of the form $\{x : f(x) < a\}$ is convex for $a \in \mathbb{R}$). However, these two properties are equivalent due to the linearity of $\mathcal{F}$ along rays.

Now consider Problem 3. By Theorem 1, there is a solution of the form

$$c(t) = \frac{1}{2\pi}[(1 - \cos(2\pi t))u + \sin(2\pi t)v] \tag{3.3}$$

for a pair of orthonormal vectors u and v in $\mathbb{R}^n$. For c given by (3.3), direct computation shows that $h(c) = \frac{1}{4\pi}\beta(u, v)$; thus

$$\mathcal{F}[h(c)] = \frac{\mathcal{F}[\beta(u, v)]}{4\pi}. \tag{3.4}$$

So to solve Problem 3 it suffices to find the pair (u, v) that maximizes $\mathcal{F}[\beta(u, v)]$.

In the following, we will outline the procedures to determine the simple circle solutions in the cases when $\mathcal{F}$ is linear and when $\mathcal{F}$ is a norm.

3.1. Solution circles when $\mathcal{F}$ is linear.

In this section, we briefly outline the procedure for finding the solution circles when $\mathcal{F}$ is a linear function. Such a procedure is well known in the literature and we include it here for completeness. Suppose that $\mathcal{F}(h) = \rho^T h$ is linear for some constant $\rho \in \mathbb{R}^k$. Then

$$\mathcal{F}[\beta(u, v)] = \rho^T \beta(u, v) = u^T Z_0 v,$$

where Z_0 is the skew-symmetric matrix defined by

$$Z_0 = \begin{bmatrix} 2\rho^T\beta_{11} & \cdots & 2\rho^T\beta_{1n} \\ \vdots & \vdots & \vdots \\ 2\rho^T\beta_{n1} & \cdots & 2\rho^T\beta_{nn} \end{bmatrix}.$$

Denote by $\sigma_1(Z_0)$ the largest singular value of Z_0. Then it is a well-known fact in linear algebra that the orthonormal u and v that maximize $u^T Z_0 v$ must be the left and right singular vectors of Z_0 corresponding to the singular value $\sigma_1(Z_0)$, i.e.,

$$Z_0 u = \sigma_1(Z_0)v, \quad Z_0 v = -\sigma_1(Z_0)u. \tag{3.5}$$

Together, u and v span a two-dimensional subspace of $\mathbb{R}^n$ invariant under Z_0. A solution to Problem 3 is then given by (3.3). A solution to Problem 2 is a scaled version of (3.3). These solutions are well known for the microorganism swimming problem, for example, in [22] where $\mathbb{R}^k = \mathbb{R}^3$ is the space of translations of the microorganism and ρ is aligned with the positive z-axis.

Note that (u, v) satisfying (3.5) is in general not unique for two reasons: Z_0 could have multiple singular values equal to $\sigma_1(Z_0)$; and even if this is not the case, a rotation of u and v within the plane they span will yield a new orthonormal pair satisfying (3.5). As a result, any simple circle of unit energy through the origin and contained in an invariant plane of Z_0 corresponding to $\sigma_1(Z_0)$ will solve Problem 3.

3.2. Solution circles when $\mathcal{F}$ is a norm.

Suppose that $\mathcal{F} = \|\cdot\|$ is a norm on $\mathbb{R}^k$. Finding the optimal circle is considerably more difficult in this case. In this section, we will propose a novel procedure, Algorithm 1, to determine the orthonormal pair (u, v) that maximizes $\mathcal{F}[\beta(u, v)] = \|\beta(u, v)\|$. A solution to Problem 3 is then given by a simple circle contained in the plane spanned by u and v.

First of all, for each $p = 1, \ldots, k$, define Z^p as the skew-symmetric matrix

$$Z^p \triangleq \begin{bmatrix} 2\beta_{11}^p & \cdots & 2\beta_{1n}^p \\ \vdots & \vdots & \vdots \\ 2\beta_{n1}^p & \cdots & 2\beta_{nn}^p \end{bmatrix},$$

where β_{ij}^p is the pth component of $\beta_{ij} \in \mathbb{R}^k$, $1 \le i, j \le n$. Let $\sigma_1(Z^p)$ be the largest singular value of Z^p, and let (u_p, v_p) be a pair of left and right singular vectors of Z^p corresponding to $\sigma_1(Z^p)$.

The solution is simple when $\|\cdot\|$ is the L^1- or L^∞-norm. So we will simply point out the results. If $\mathcal{F}$ is the L^∞-norm, the pair (u, v) that maximizes $\|\beta(u, v)\|$ is the pair (u_p, v_p) for a p with the largest $\sigma_1(Z^p)$. If $\mathcal{F}$ is the L^1-norm, then define $\mathcal{A} \triangleq \{\pm Z^1 \pm Z^2 \pm \cdots \pm Z^k\}$, and choose a $Z \in \mathcal{A}$ with the largest $\sigma_1(Z)$. The pair (u, v) that maximizes $\|\beta(u, v)\|$ is then given by a pair of left and right singular vectors of Z corresponding to $\sigma_1(Z)$.

In the rest of this section we focus on the more interesting case where $\|\cdot\|$ is the L^2-norm. In this case, since β is antisymmetric, the pair (u, v) maximizing $\|\beta(u, v)\|$ subject to $\|u\| = \|v\| = 1$ will automatically be orthogonal. So we might as well drop the orthogonality constraint. Write

$$\|\beta(u, v)\|^2 = \sum_{p=1}^{k} (u^T Z^p v)^2 = u^T \left(\sum_{p=1}^{k} Z^p v v^T (Z^p)^T \right) u = v^T \left(\sum_{p=1}^{k} (Z^p)^T u u^T Z^p \right) v.$$

Therefore, to maximize $\|\beta(u, v)\|$ under the constraint that $\|u\| = \|v\| = 1$, we need the following conditions:

(i) u is an eigenvector of $\sum_{p=1}^{k} Z^p v v^T (Z^p)^T$ for its largest eigenvalue;

(ii) v is an eigenvector of $\sum_{p=1}^{k} (Z^p)^T u u^T Z^p$ for its largest eigenvalue.

These two conditions hint at the following iterative algorithm.

ALGORITHM 1. *Choose some initial u and v in $\mathbb{R}^n$ such that $\|u\| = \|v\| = 1$.*

1. *Let u be a unit eigenvector of $\sum_{p=1}^{k} Z^p v v^T (Z^p)^T$ for its largest eigenvalue.*
2. *Let v be a unit eigenvector of $\sum_{p=1}^{k} (Z^p)^T u u^T Z^p$ for its largest eigenvalue.*
3. *Repeat steps* 1 *and* 2 *until some convergence criteria is satisfied, for example, when the changes in u, v in consecutive steps are below a given threshold.*

The value of $\|\beta(u, v)\|$ increases with each iteration, and, barring the occurrence of cycles, u and v will converge to a pair satisfying conditions (i) and (ii). However, as these are only necessary conditions, the convergence property to the global solutions is still an issue to be resolved.

Remark 5. Bounds on $\max\{\|\beta(u, v)\| : \|u\| = \|v\| = 1\}$ can be obtained as follows. For each $p = 1, \ldots, k$, let the column vectors of Z^p from left to right be stacked from top to bottom into a single column vector $\mathbf{z}_p \in \mathbb{R}^{n^2}$. In addition, for $u = (u_1, \ldots, u_n)$ and $v = (v_1, \ldots, v_n)$ in $\mathbb{R}^n$, denote by $u \otimes v$ the vector $(u_1 v_1, \ldots, u_1 v_n, \ldots, u_n v_1, \ldots, u_n v_n) \in \mathbb{R}^{n^2}$. Then $\mathbf{z}_p^T (u \otimes v) = u^T Z^p v$, and

$$\beta(u, v) = \begin{bmatrix} u^T Z^1 v \\ \vdots \\ u^T Z^k v \end{bmatrix} = \begin{bmatrix} \mathbf{z}_1^T (u \otimes v) \\ \vdots \\ \mathbf{z}_k^T (u \otimes v) \end{bmatrix} = \mathbf{Z}(u \otimes v),$$

where $\mathbf{Z} \in \mathbb{R}^{k \times n^2}$ is defined by $\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1 & \cdots & \mathbf{z}_k \end{bmatrix}^T$. Since $u \otimes v$ is a unit vector in $\mathbb{R}^{n^2}$ for unit u, v, we have

$$\|\beta(u,v)\| = \|\mathbf{Z}(u \otimes v)\| \leq \sigma_1(\mathbf{Z}). \tag{3.6}$$

In general, $\{u \otimes v : \|u\| = \|v\| = 1\}$ is a proper subset of the unit sphere in $\mathbb{R}^{n^2}$. So the bound (3.6) is not strict. By rearranging Z^p in different ways, we can obtain other bounds similar to (3.6). See [9] for more on the singular value decomposition of multilinear tensors.

In the case when the base space has dimension three, i.e., when $n = 3$, the solution is especially simple. In fact, for each $p = 1, \ldots, k$, since $Z^p \in \mathbb{R}^{3 \times 3}$ is skew-symmetric, we can find $z_p \in \mathbb{R}^3$ such that $Z^p v = v \times z_p$ for all $v \in \mathbb{R}^3$. Here × denotes the cross product of vectors in $\mathbb{R}^3$. Therefore,

$$\beta(u,v) = \begin{bmatrix} u^T Z^1 v \\ \vdots \\ u^T Z^k v \end{bmatrix} = \begin{bmatrix} u^T (v \times z_1) \\ \vdots \\ u^T (v \times z_k) \end{bmatrix} = \begin{bmatrix} z_1^T (u \times v) \\ \vdots \\ z_k^T (u \times v) \end{bmatrix} = \mathbf{Z}(u \times v), \tag{3.7}$$

where $\mathbf{Z} \in \mathbb{R}^{k \times 3}$ is defined by $\mathbf{Z} \triangleq \begin{bmatrix} z_1 & \cdots & z_k \end{bmatrix}^T$. Note that the set of $u \times v$ for unit u and v is exactly the unit ball in $\mathbb{R}^3$. Therefore,

$$\|\beta(u,v)\| = \|\mathbf{Z}(u \times v)\| \leq \sigma_1(\mathbf{Z}),$$

with exact equality achieved by any orthonormal pair u and v with $u \times v = w$, where $w \in \mathbb{R}^3$ is a unit right singular vector of $\mathbf{Z}$ corresponding to $\sigma_1(\mathbf{Z})$.

4. Higher rank case. Solving Problems 2 and 3 in the higher rank case is much more difficult than in the rank two case, as analytic characterization of solutions is in general not available. In this section, however, we shall study a special class of higher rank problems for which analytical characterization is possible. The result will then be applied in section 5 to the linked-mass system in section 1.1.

Consider the following codimension one distribution $\mathcal{H}$ on $\mathbb{R}^3$ with base space $\mathbb{R}^2$. The forms α and β specifying $\mathcal{H}$ as in (2.4) and (2.6) are, respectively,

$$\alpha = x_1^r dx_2, \quad \beta = r x_1^{r-1}\, dx_1 \wedge dx_2, \tag{4.1}$$

for some integer $r \geq 2$. When $r = 2$, this distribution is called the Martinet distribution and has been well studied in the subriemmanian geometry and optimal control literature (see, e.g., [1, 15]). By Lemma 2, the rank of $\mathcal{H}$ at the origin is $r + 1$. Suppose that the metric $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is obtained by lifting the standard metric on $\mathbb{R}^2$, and that $\mathcal{F} : \mathbb{R} \to \mathbb{R}$ is the identity map. So Assumption 1 in section 2.4 is satisfied.

For a loop c in $\mathbb{R}^2$ based at 0 enclosing a surface S, $h(c) = \int_c x_1^r dx_2 = \int_S r x_1^{r-1}\, dx_1 \wedge dx_2$. Problem 2 then reduces to the following:

$$\text{Find } c \text{ with } \int_c x_1^r dx_2 = 1 \text{ that minimize } E(c). \tag{4.2}$$

Lemma 6. *There is a solution c to problem* (4.2) *such that*

1. *c is contained exclusively in the closed right half plane;*
2. *c has no self-crossing and encloses a convex region S;*
3. *S is symmetric with respect to the x_1-axis.*

To prove each claim one shows that for a loop c not satisfying the condition, a better loop can be obtained by properly transforming c. As an example, for claim 2 one can "flip" outside a certain segment of a nonconvex c contained strictly in its convex hull to obtain a loop with the same energy but a larger holonomy. We omit the proof here.

Let c be a solution to problem (4.2) satisfying the conditions in Lemma 6. Because of the symmetry, it suffices to study the first half of c only, which starts from the origin at time 0, follows the graph of a convex function f below the x_1-axis during $[0,1/2]$, and reaches a point $(a,0)$ with $a>0$ on the x_1-axis at time $1/2$. Such a $c=(x_1,x_2)$ must solve the following optimal control problem:

$$\text{(4.3)}\qquad \text{Minimize}\qquad \frac{1}{2}\int_0^{\frac{1}{2}}(u_1^2+u_2^2)\,dt$$

$$\text{subject to}\ \begin{cases}\dot{x}_1=u_1, & x_1(0)=0,\ x_1(\frac{1}{2})=a,\\ \dot{x}_2=u_2, & x_2(0)=0,\ x_2(\frac{1}{2})=0,\\ \dot{x}_3=x_1^r u_2, & x_3(0)=0,\ x_3(\frac{1}{2})=\frac{1}{2}.\end{cases}$$

Note that a is a parameter to be determined later so that $\dot{x}_1(\frac{1}{2})=0$; thus the two halves of c can be pieced together smoothly. The boundary condition $x_3(\frac{1}{2})=\frac{1}{2}$ is imposed to ensure that $\int_c x_1^2dx_2=1$.

Define the Hamiltonian as follows:

$$H(\lambda_1,\lambda_2,\lambda_3,x_1,x_2,x_3)=\frac{u_1^2+u_2^2}{2}+\lambda_1u_1+\lambda_2u_2+\lambda_3x_1^ru_2.$$

By the maximum principle [19], u_1,u_2 for the optimal solution can be determined as

$$\text{(4.4)}\qquad u_1=\operatorname{argmin}_{u_1}H=-\lambda_1,$$

$$\text{(4.5)}\qquad u_2=\operatorname{argmin}_{u_2}H=-\lambda_2-\lambda_3x_1^r,$$

while λ_i, $i=1,2,3$, satisfy

$$\dot{\lambda}_1=-\frac{\partial H}{\partial x_1}=-r\lambda_3x_1^{r-1}u_2,\qquad \dot{\lambda}_2=-\frac{\partial H}{\partial x_2}=0,\qquad \dot{\lambda}_3=-\frac{\partial H}{\partial x_3}=0.$$

Thus λ_2 and λ_3 are constant. Their signs can be determined as $\lambda_2\ge 0$ and $\lambda_3<0$. In fact, at time $t=0$ in (4.5), we have $x_1=0$; hence $-\lambda_2=u_2(0)=\dot{x}_2(0)\le 0$ by our assumption that the curve f is below the x_1-axis during $[0,1/2]$, i.e., $\lambda_2\ge 0$. Denote by $\tau\in(0,\frac{1}{2})$ the time when x_2 achieves its minimum during $[0,\frac{1}{2}]$. Thus, $\dot{x}_2(\tau)=-\lambda_2-\lambda_3x_1^2(\tau)=0$, which is possible only if $\lambda_3\le 0$. Moreover, $\lambda_3\ne 0$, for otherwise $\dot{x}_2=u_2=-\lambda_2$ is constant zero, an obvious contradiction.

Note that $\ddot{x}_1=\dot{u}_1=-\dot{\lambda}_1=r\lambda_3x_1^{r-1}u_2=-r\lambda_3x_1^{r-1}(\lambda_2+\lambda_3x_1^r)$. Hence,

$$d(\dot{x}_1^2)=2\ddot{x}_1dx_1=-2r\lambda_3x_1^{r-1}(\lambda_2+\lambda_3x_1^r)\dot{x}_1=-d(\lambda_2+\lambda_3x_1^r)^2.$$

The integrability of the above equation is a key result that can greatly reduce the complexity of solving the optimal control problem (4.3). After integration, we obtain

$$\dot{x}_1^2=\lambda_3^2C^2-(\lambda_2+\lambda_3x_1^r)^2$$

for some constant $C>0$. Therefore,

$$\text{(4.6)}\qquad \dot{x}_1=-\lambda_3\sqrt{C^2-(\lambda_2/\lambda_3+x_1^r)^2}.$$

Note that $|C| \geq \lambda_2/\lambda_3$, for otherwise $\dot{x}_1$ is not defined at time 0. Since $x_1 = a$ and $\dot{x}_1 = 0$ at time $t = \frac{1}{2}$, a can be determined as

$$a = (C - \lambda_2/\lambda_3)^{1/r}. \tag{4.7}$$

The graph of the function f that c follows during $[0, \frac{1}{2}]$ can be derived directly. Dividing $\dot{x}_2 = u_2 = -(\lambda_2 + \lambda_3 x_1^r)$ by (4.6), we have

$$\frac{dx_2}{dx_1} = \frac{\lambda_2/\lambda_3 + x_1^r}{\sqrt{C^2 - (\lambda_2/\lambda_3 + x_1^r)^2}}, \quad 0 \leq x_1 \leq a. \tag{4.8}$$

Integrating the above equation with respect to x_1 will yield x_2 as a function of $x_1 \in [0, \frac{1}{2}]$, namely, the graph of the function f. It remains to determine the unknown parameters λ_2/λ_3 and C in (4.7). The boundary conditions $x_2(\frac{1}{2}) = 0$ and $x_3(\frac{1}{2}) = \frac{1}{2}$ imply, respectively, that

$$\int_0^a \frac{\lambda_2/\lambda_3 + x_1^r}{\sqrt{C^2 - (\lambda_2/\lambda_3 + x_1^r)^2}}\, dx_1 = 0, \tag{4.9}$$

$$\int_0^a \frac{x_1^r(\lambda_2/\lambda_3 + x_1^r)}{\sqrt{C^2 - (\lambda_2/\lambda_3 + x_1^r)^2}}\, dx_1 = \frac{1}{2}. \tag{4.10}$$

The procedures to determine λ_2/λ_3 and C satisfying the above conditions are as follows:

1. Choose any fixed λ_2/λ_3, say, $\lambda_2/\lambda_3 = \kappa_0 < 0$.
2. Find C so that (4.9) is satisfied, say, $C = C_0$. Note that a in (4.9) is determined by (4.7).
3. Use κ_0 and C_0 in (4.7) to compute an a, say, $a = a_0$.
4. Use κ_0, C_0, and a_0 in (4.8) to integrate for a function $x_2 = g(x_1)$ on $[0, a_0]$. The function g obtained so far is in general not the desired function f, since constraint (4.10) may not be satisfied. However, f can be obtained from g by a proper scaling. In fact, define

$$\mu = \left[2 \int_0^{a_0} \frac{x_1^r(\kappa_0 + x_1^r)}{\sqrt{C_0^2 - (\kappa_0 + x_1^r)^2}}\, dx_1\right]^{1/(r+1)}.$$

5. Define a function f by $f(x_1) = g(\mu x_1)/\mu$ for $x_1 \in [0, a_0/\mu]$.

It can be verified that the obtained f satisfies (4.9) and (4.10) for $\lambda_2/\lambda_3 = \kappa_0\mu^{-r}$ and $C = C_0\mu^{-r}$ and is indeed the desired function whose graph c follows during $[0, \frac{1}{2}]$. If one is interested only in the shape, not the scale, of the solution to problem (4.2), then the last step can be skipped.

Remark 6. The time parameterization of c is recovered using the fact that c has constant speed along the graph of f on $[0, \frac{1}{2}]$. Alternatively, by (4.6),

$$t = \Phi(x_1) \triangleq \int_0^{x_1} \frac{dx_1}{-\lambda_3\sqrt{C^2 - (\lambda_2/\lambda_3 + x_1^r)^2}}, \quad 0 \leq x_1 \leq a.$$

Note that $\Phi(x_1)$ is a strictly increasing function satisfying $\Phi(a) = \frac{1}{2}$. Once the function $\Phi(x_1)$ is determined, x_1 is determined as $x_1 = \Phi^{-1}(t)$.

Figure 4.1 plots the solution loops to problem (4.3) obtained from the above procedure for the case $r = 2, 4, 10$. In particular, in the Martinet distribution case

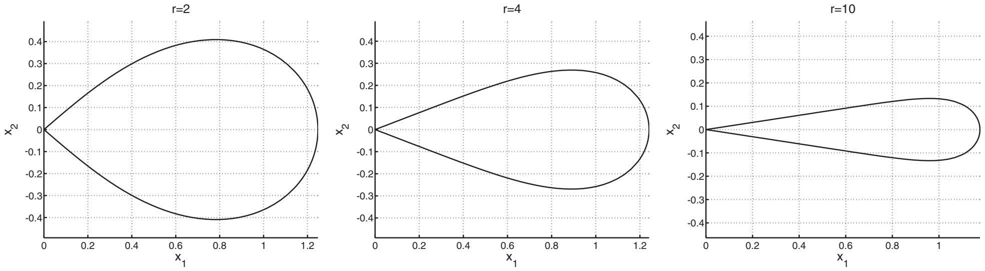


FIG. 4.1. *Solution loops to problem* (4.3) *when* $r = 2$ *(left),* $r = 4$ *(middle),* $r = 10$ *(right).*

($r = 2$), the solution loop is part of the trajectory of a charged particle moving on $\mathbb{R}^2$ in a magnetic field $\mathbf{B}$ with linear components and direction perpendicular to the plane (see the more complete plot of the motion in Figure 2.2 of Alfven's book [2, p. 15]). The trajectory is called the *grad* $\mathbf{B}$ *drift* since it exhibits an overall drift orthogonal to the direction grad$|\mathbf{B}|$. Indeed, much more has been known for the Martinet distribution: for example, its unit subriemannian sphere around the origin has been characterized and plotted using elliptic integrals in [1]. That the equation governing the solution loops is integrable in the $r = 2$ case has also been pointed out previously in [23].

5. A three-segment linked-mass system.

We now return to the motivating example in section 1 and consider the case with $n = 2$. So the linked-mass system consists of four nodes, and three links whose orientations are given by the angles θ_i, $i = 1, 2, 3$. The configuration space is $\mathbb{T}^3$, with a riemannian metric given by (1.6), where

$$\Delta = (\Delta_{ij})_{i,j=1}^3 = \frac{1}{4}\begin{bmatrix} 3 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 3 \end{bmatrix}.$$

By (1.8), the codimension one distribution $\mathcal{H}$ is the kernel of the one-form

$$\omega = \sum_{i,j=1}^{3} \Delta_{ij} \cos(\theta_i - \theta_j) d\theta_j.$$

Suppose now that the linked-mass system is at the initial configuration q corresponding to $\theta_1 = \theta_2 = \theta_3 = 0$; i.e., the three segments of the system are all aligned in the positive horizontal direction. To compute the rank at q, we perform the following coordinate transformation in a neighborhood of q:

$$(5.1) \qquad \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \end{bmatrix} = \begin{bmatrix} -\frac{\sqrt{5}}{3} & \frac{2\sqrt{5}}{3} & -\frac{\sqrt{5}}{3} \\ -1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}, \quad \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} = \begin{bmatrix} -\frac{\sqrt{5}}{10} & -\frac{1}{2} & \frac{1}{3} \\ \frac{\sqrt{5}}{5} & 0 & \frac{1}{3} \\ -\frac{\sqrt{5}}{10} & \frac{1}{2} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \end{bmatrix}.$$

The choice of such a transformation serves several purposes. First, $\phi_3 = \theta_1 + \theta_2 + \theta_3$ is the direction along the fibers of $\mathbb{T}^3$ under the action of $\mathbb{T}$ as described in section 1. Second, the plane Π spanned by the ϕ_1 and ϕ_2 axes is transversal to the ϕ_3 axis, and hence can be regarded as the shape space, at least locally around the origin. Third, the projection $d\pi : \mathcal{H}_q \to T_{(0,0)}\Pi$ is an isometry if Π is equipped with the canonical

Euclidean metric with respect to the coordinates (ϕ_1, ϕ_2). So no further change of coordinates within Π is needed.

In the new coordinates, q corresponds to the origin $\phi_1 = \phi_2 = \phi_3 = 0$, and

$$\begin{aligned}
&4\omega \\
&\quad = 3d\theta_1 + 2C_{12}d\theta_2 + C_{13}d\theta_3 + 2C_{12}d\theta_1 + 4d\theta_2 + 2C_{23}d\theta_3 + C_{13}d\theta_1 + 2C_{23}d\theta_2 + 3d\theta_3 \\
&\quad = \frac{\sqrt{5}}{5}(1 + C_{12} + C_{23} - C_{13})d\phi_1 + (C_{23} - C_{12})d\phi_2 + \frac{2}{3}(5 + 2C_{12} + 2C_{23} + C_{13})d\phi_3,
\end{aligned}$$

where C_{12}, C_{23}, C_{13} are defined by

$$\begin{aligned}
C_{12} &\triangleq \cos(\theta_1 - \theta_2) = \cos\left(\frac{3\sqrt{5}}{10}\phi_1 + \frac{1}{2}\phi_2\right), \\
C_{23} &\triangleq \cos(\theta_2 - \theta_3) = \cos\left(\frac{3\sqrt{5}}{10}\phi_1 - \frac{1}{2}\phi_2\right), \\
C_{13} &\triangleq \cos(\theta_1 - \theta_3) = \cos\phi_2.
\end{aligned}$$

From the above equations, the kernel of ω is the same as the kernel of

$$\Theta = -\frac{3(C_{12} - C_{23})}{2(5 + 2C_{12} + 2C_{23} + C_{13})}d\phi_1 + \frac{3(1 + C_{12} + C_{23} - C_{13})}{2\sqrt{5}(5 + 2C_{12} + 2C_{23} + C_{13})}d\phi_2 + d\phi_3,$$

which is of the standard form (2.1). Note that C_{12}, C_{23}, C_{13} are independent of ϕ_3, as the distribution $\mathcal{H}$ is invariant to the bundle structure $\pi : (\phi_1, \phi_2, \phi_3) \mapsto (\phi_1, \phi_2)$.

The form α defined in (2.4) is given by

$$\begin{aligned}
\alpha &= -\frac{3(C_{12} - C_{23})}{2(5 + 2C_{12} + 2C_{23} + C_{13})}d\phi_1 + \frac{3(1 + C_{12} + C_{23} - C_{13})}{2\sqrt{5}(5 + 2C_{12} + 2C_{23} + C_{13})}d\phi_2 \\
&= \frac{3\sin(\frac{3\sqrt{5}}{10}\phi_1)\sin(\frac{1}{2}\phi_2)}{5 + 4\cos(\frac{3\sqrt{5}}{10}\phi_1)\cos(\frac{1}{2}\phi_2) + \cos\phi_2}d\phi_1 \\
&\qquad + \frac{3(1 + 2\cos(\frac{3\sqrt{5}}{10}\phi_1)\cos(\frac{1}{2}\phi_2) - \cos\phi_2)}{2\sqrt{5}(5 + 4\cos(\frac{3\sqrt{5}}{10}\phi_1)\cos(\frac{1}{2}\phi_2) + \cos\phi_2)}d\phi_2,
\end{aligned}$$

and $\beta = d\alpha = f(\phi_1, \phi_2)d\phi_1 \wedge d\phi_2$, where $f(\phi_1, \phi_2)$ is given by

$$\begin{aligned}
&f(\phi_1, \phi_2) \\
&\quad = \frac{-3\sin(\frac{3\sqrt{5}}{10}\phi_1)}{5(5 + 4\cos(\frac{3\sqrt{5}}{10}\phi_1)\cos(\frac{1}{2}\phi_2) + \cos\phi_2)} \\
&\quad \times \left\{4\cos\left(\frac{1}{2}\phi_2\right) + \frac{\cos(\frac{3\sqrt{5}}{10}\phi_1)[10\sin^2(\frac{1}{2}\phi_2) - 6\cos^2(\frac{1}{2}\phi_2)] + 4\sin^2(\frac{1}{2}\phi_2)\cos(\frac{1}{2}\phi_2)}{5 + 4\cos(\frac{3\sqrt{5}}{10}\phi_1)\cos(\frac{1}{2}\phi_2) + \cos\phi_2}\right\}.
\end{aligned} \tag{5.2}$$

One can verify that

$$f(0,0) = 0, \quad \frac{\partial f}{\partial\phi_1}(0,0) = -\frac{51}{250} \neq 0, \quad \frac{\partial f}{\partial\phi_2}(0,0) = 0.$$

As a result of Lemma 2, the rank at q is three.

Suppose that the test function $\mathcal{F} : \mathbb{R} \to \mathbb{R}$ is the identity map. To find the asymptotically optimal loop c based at q in the plane Π solving Problem 1, we can replace β by its first order approximate $-\frac{51}{250}\phi_1 d\phi_1 \wedge d\phi_2$, which is exactly of the form (4.1) with $r = 2$. Thus the results in section 4 can be applied directly here.

In particular, the optimal loop c is computed in section 4 and plotted in Figure 4.1 with coordinates $x_1 = \phi_1$ and $x_2 = \phi_2$. Horizontally lifting c in Π to a curve γ based at q in the (ϕ_1, ϕ_2, ϕ_3) coordinates, transforming γ back to the $(\theta_1, \theta_2, \theta_3)$ coordinates using transformation (5.1), and finally, using (1.4) and (1.5), we obtain an asymptotically most efficient motion for the linked-mass system starting from the initially aligned position. Figure 5.1 shows the snapshots of the motions obtained numerically at equally spaced time instances. Note that a relatively large scale of c is chosen in the plots to make this asymptotic motion more obvious.

Remark 7. By (5.2), in a neighborhood of the origin in the (ϕ_1, ϕ_2) coordinates, $\beta = 0$ if and only if $\phi_1 = 0$, i.e., if and only if $\theta_1 + \theta_3 - 2\theta_2 = 0$. The rank is three at points satisfying this condition and is two otherwise. As a result, when the system starts from a shape close to the aligned one, in the asymptotic sense it is more difficult to turn if its initial position is such that $\theta_1 + \theta_3 - 2\theta_2 = 0$.

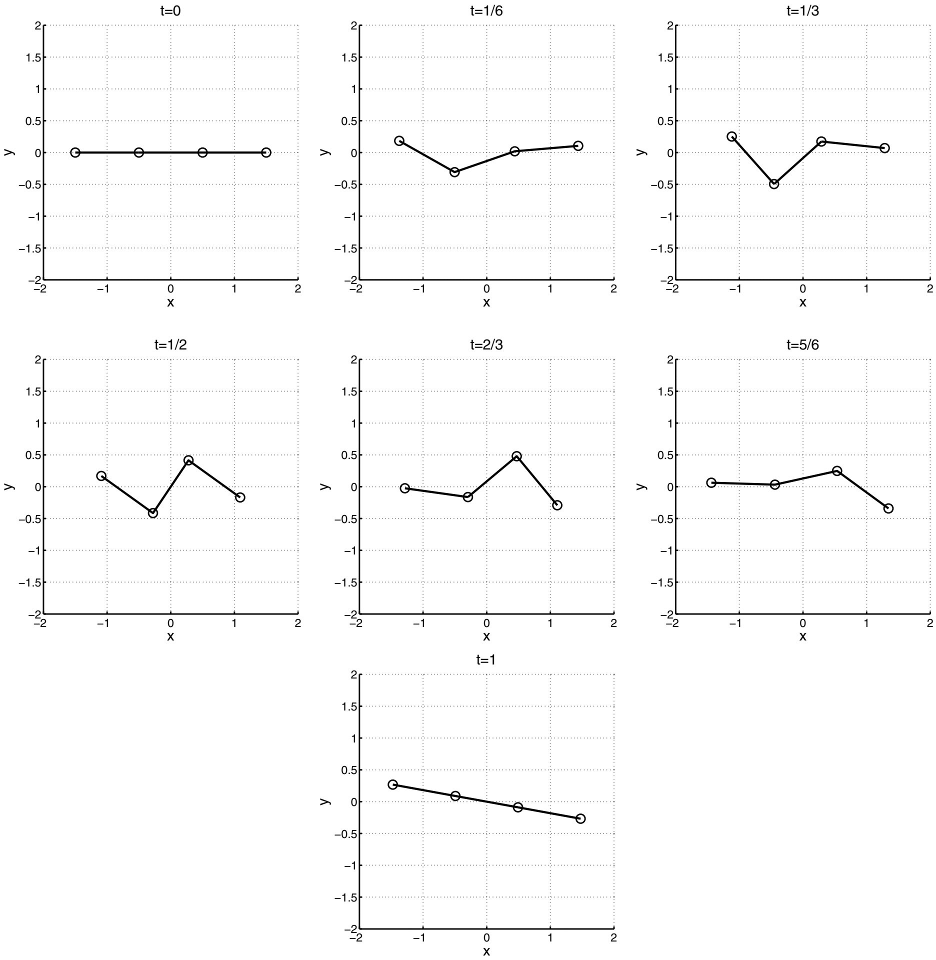


FIG. 5.1. *Snapshots of the linked-mass system turning.*

6. Extension to principal bundles. Due to their asymptotic nature, our results can be easily generalized to more complicated spaces such as principal bundles, as is described briefly in the following. For a Lie group G with Lie algebra $\mathfrak{g}$, a principal G-bundle $\pi : Q \to M$ is a fiber bundle whose structural group G acts freely and transitively on each fiber from the right. So each fiber is a copy of G, and the vertical space $\mathcal{V}_q = \mathrm{im}(\sigma_q)$ at each $q \in Q$ can be identified with $\mathfrak{g}$ via the map $\sigma_q : \xi \in \mathfrak{g} \mapsto q \cdot \xi \in T_qQ$. A connection form Θ on Q is a $\mathfrak{g}$-valued one-form with $\ker \Theta_q \oplus \mathcal{V}_q = T_qQ$ and $\Theta_q \circ \sigma_q = \mathrm{id}_{\mathfrak{g}}$ for all $q \in Q$. Thus $\mathcal{H} \triangleq \ker \Theta$ defines a horizontal distribution on Q invariant under the action of G. The holonomy $h(c)$ of a loop c in M based at m can be identified as an element of G and is determined up to a conjugacy class in G when varying the base point q of the horizontal lift [18].

Suppose that $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ is a subriemannian metric on $\mathcal{H}$ invariant under the action of G. Such $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ is obtained by lifting a riemannian metric $\langle\cdot,\cdot\rangle_M$ on M. For a loop c in M based at m, we can define the rank $r_q(m) \in \mathbb{N}$ and the asymptotic holonomy $\eta_q(c) \in \mathfrak{g}$ such that $h(c_\epsilon) \sim \eta_q(c)[L(c_\epsilon)]^{r_q(m)}$ as $\epsilon \to 0$. Here the scaled loop c_ϵ is defined by identifying M locally around m with an open subset of $\mathbb{R}^n$ via a coordinate map, for example, the inverse of the exponential map $\exp : T_mM \to M$. It is easy to see that both $r_q(m)$ and $\eta_q(c)$ are independent of the choice of the coordinate maps; thus they are well defined. A test function is a map $\mathcal{F} : \mathfrak{g} \to \mathbb{R}$. For example, $\mathcal{F}$ can be the inertial tensor $\mathcal{F}(\xi) = \langle\sigma_q(\xi), \sigma_q(\xi)\rangle_Q$ for all $\xi \in \mathfrak{g}$ for some metric $\langle\cdot,\cdot\rangle_Q$ on Q that restricts to $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ on $\mathcal{H}$.

With the above notions, we can define the asymptotic holonomic efficiency problem on principal bundles. In a neighborhood of q, Q and $\mathbb{R}^n \oplus \mathfrak{g}$ are the same in terms of computing asymptotic holonomy. For example, when the distribution is of rank two at a point $m \in M$, it can be proved similarly as in Theorem 1 that the optimal loop c is a circle in a plane spanned by two tangent vectors u_m, v_m in T_mM. Finding the optimal circle then becomes the problem of finding these two tangent vectors u_m and v_m that maximizes a sectional curvature-like term $\mathcal{F}[\beta(u_m, v_m)]$ as in (3.4). In the higher rank case, however, finding the optimal solutions is a much more challenging problem.

Finally, in terms of infinitesimal deformations, the general case when G is nonabelian looks exactly the same locally as the abelian case. Therefore, our results can be extended to the general nonabelian case without added difficulty [21].

Appendix. Computation of the metric on $\mathbb{T}^{n+1}$. Plugging (1.4) into (1.5), we have

$$q_i = \frac{1}{n+2}\sum_{j=1}^{i-1} j(\cos\theta_j, \sin\theta_j)^T - \frac{1}{n+2}\sum_{j=i}^{n+1}(n+2-j)(\cos\theta_j, \sin\theta_j)^T. \tag{A.1}$$

As a result, the tangent vector $\frac{\partial}{\partial\theta_l}$ at each point of $\mathbb{T}^{n+1}$ is pushed forward by the embedding defined by (1.4) and (1.5) to a tangent vector in $\mathbb{R}^{2n+4}$:

$$\frac{\partial q}{\partial\theta_l} = \Bigg[\underbrace{-\frac{n+2-l}{n+2}(-\sin\theta_l, \cos\theta_l)}_{\text{repeated } l \text{ times}} \quad \underbrace{\frac{l}{n+2}(-\sin\theta_l, \cos\theta_l)}_{\text{repeated } n+2-l \text{ times}}\Bigg]^T \in \mathbb{R}^{2n+4}. \tag{A.2}$$

The metric on $\mathbb{T}^{n+1}$ can be derived from the standard metric on $\mathbb{R}^{2n+4}$ as $\langle\frac{\partial}{\partial\theta_i}, \frac{\partial}{\partial\theta_j}\rangle = \left(\frac{\partial q}{\partial\theta_i}\right)^T \cdot \left(\frac{\partial q}{\partial\theta_j}\right)$. Using (A.2), we can easily verify that $\langle\frac{\partial}{\partial\theta_i}, \frac{\partial}{\partial\theta_j}\rangle = \frac{i(n+2-j)}{n+2}\cos(\theta_i - \theta_j)$ if $i < j$ and $\langle\frac{\partial}{\partial\theta_i}, \frac{\partial}{\partial\theta_j}\rangle = \frac{(n+2-i)j}{n+2}\cos(\theta_i - \theta_j)$ if $i \geq j$.

Acknowledgments. The authors would like to thank Richard Montgomery for pointing out some of the references and Alan Weinstein for helpful discussions. We are also grateful to the anonymous reviewers for their constructive comments and for pointing out the many links to the existing literature.

REFERENCES

[1] A. A. Agrachev, B. Bonnard, M. Chyba, and I. Kupka, *Sub-Riemannian sphere in Martinet flat case*, ESAIM Control Optim. Calc. Var., 2 (1997), pp. 377–448.

[2] H. Alfven, *Cosmical Electrodynamics*, Oxford University Press, Oxford, UK, 1950.

[3] J. B. Baillieul, *Geometric methods for nonlinear optimal control problems*, J. Optim. Theory Appl., 25 (1978), pp. 517–548.

[4] J. R. Blake, *Self propulsion due to oscillations on the surface of a cylinder at low Reynolds numbers*, Bull. Austral. Math. Soc., 5 (1971), pp. 255–264.

[5] A. M. Bloch, *Nonholonomic Mechanics and Control*, Springer-Verlag, New York, 2003.

[6] R. W. Brockett, *Control theory and singular Riemannian geometry*, in New Directions in Applied Mathematics, P. J. Hilton and G. S. Young, eds., Springer-Verlag, New York, 1982, pp. 11–27.

[7] S. Childress, *Mechanics of Swimming and Flying*, Cambridge University Press, Cambridge, UK, 1981.

[8] J. Cortes, S. Martnez, J. P. Ostrowski, and K. A. McIsaac, *Optimal gaits for dynamic robotic locomotion*, Internat. J. Robotics Res., 20 (2001), pp. 707–728.

[9] L. de Lathauwer, B. de Moor, and J. Vandewalle, *A multilinear singular value decomposition*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1253–1278.

[10] A. Guichardet, *On rotation and vibration motions of molecules*, Ann. Inst. H. Poincaré Phys. Théor., 40 (1984), pp. 329–342.

[11] J. Hu, S. Simic, and S. Sastry, *How should a snake turn on ice: A case study of the asymptotic isoholonomic problem*, in Proceedings of the IEEE International Conference on Decision and Control, Vol. 3, Maui, HI, 2003, pp. 2908–2913.

[12] T. R. Kane and M. P. Scher, *A dynamical explanation of the falling cat phenomenon*, Internat. J. Solids Structures, 5 (1969), pp. 663–670.

[13] J. Koiller and J. Delgado, *On efficiency calculations for nonholonomic locomotion problems: An application to microswimming*, Rep. Math. Phys., 42 (1998), pp. 165–183.

[14] J. M. Lighthill, *On the squirming motion of nearly spherical deformable bodies through liquids at very small Reynolds number*, Comm. Pure Appl. Math., 5 (1952), pp. 109–118.

[15] W. Liu and H. J. Sussmann, *Shortest paths for sub-Riemannian metrics on rank-two distributions*, Mem. Amer. Math. Soc., 118 (1995), no. 564.

[16] J. E. Marsden and T. S. Ratiu, *Introduction to Mechanics and Symmetry*, 2nd ed., Springer-Verlag, New York, 1999.

[17] R. Montgomery, *The isoholonomic problem and some applications*, Comm. Math. Phys., 128 (1990), pp. 565–592.

[18] R. Montgomery, *A Tour of Sub-Riemannian Geometries, Their Geodesics and Applications*, AMS, Providence, RI, 2002.

[19] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mischenko, *Mathematical Theory of Optimal Processes*, Interscience Publishers, London, New York, 1962.

[20] S. Sastry and R. Montgomery, *The structure of optimal controls for a steering problem*, in Proceedings of the 2nd IFAC Symposium on Nonlinear Control System Design (NOLCOS'92), Bordeaux, France, 1992, pp. 385–390.

[21] A. Shapere and F. Wilczek, *Self-propulsion at low Reynolds number*, Phys. Rev. Lett., 58 (1987), pp. 2051–2054.

[22] A. Shapere and F. Wilczek, *Efficiencies of self-propulsion at low Reynolds number*, J. Fluid Mech., 198 (1989), pp. 587–599.

[23] R. Yang, *Nonholonomic Geometry, Mechanics and Control*, Ph.D. thesis, University of Maryland, College Park, MD, 1992.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 2, pp. 1134–1154

© 2009 Society for Industrial and Applied Mathematics

FINITE HORIZON OPTIMAL INVESTMENT AND CONSUMPTION WITH TRANSACTION COSTS[*]

MIN DAI[†], LISHANG JIANG[‡], PEIFAN LI[†], AND FAHUAI YI[§]

Abstract. This paper concerns continuous-time optimal investment and the consumption decision of a constant relative risk aversion (CRRA) investor who faces proportional transaction costs and a finite time horizon. In the no-consumption case, it has been studied by Liu and Loewenstein [*Review of Financial Studies*, 15 (2002), pp. 805–835] and Dai and Yi [*J. Differential Equations*, 246 (2009), pp. 1445–1469]. Mathematically, it is a singular stochastic control problem whose value function satisfies a parabolic variational inequality with gradient constraints. The problem gives rise to two free boundaries which stand for the optimal buying and selling strategies, respectively. We present an analytical approach to analyze the behaviors of free boundaries. The regularity of the value function is studied as well. Our approach is essentially based on the connection between singular control and optimal stopping, which is first revealed in the present problem.

Key words. optimal investment and consumption, transaction costs, finite horizon, free boundaries, variational inequality, gradient constraints, singular stochastic control

AMS subject classifications. 35B37, 35K85, 47J20, 49J20, 93E20

DOI. 10.1137/070703685

1. Introduction.

This paper concerns continuous-time optimal investment and the consumption decision of a constant relative risk aversion (CRRA) investor who faces a finite horizon and proportional transaction costs. In the absence of transaction costs, Merton (1971) has shown that the optimal strategy is to keep a constant fraction of total wealth in each asset and to consume at a rate proportional to wealth. Such a strategy leads to incessant trading, which is impracticable in a real market with transaction costs.

Magill and Constantinides (1976) introduced proportional transaction costs to Merton's model. They provided a fundamental insight that there exists a no-trading region and that trading only takes place along the boundary of the no-trading region. Davis and Norman (1990) first formulated the problem as a free boundary problem, where the boundary of the no-trading region is the so-called free boundary. They then studied the properties of the free boundary that reflect the optimal strategy. In terms of a viscosity solution approach, Shreve and Soner (1994) entirely characterized the behaviors of the free boundary. Akian, Menaldi, and Sulem (1996) considered an extension to the case of multiple risky assets. Janeček and Shreve (2004) presented an asymptotic expansion of the associated value function and obtained some asymptotic results on the free boundary. All of these works were confined to infinite horizon problems.

[*] Received by the editors September 25, 2007; accepted for publication (in revised form) December 1, 2008; published electronically March 11, 2009. This is an updated version of our earlier paper titled "Analysis of a Parabolic Variational Inequality with Gradient Constraints from Optimal Investment and Consumption with Transaction Costs." This research was supported by Singapore MOE AcRF grant (R-146-000-096-112), NUS RMI grant (R-146-000-117-720/646), NNSFs of China (10671075 and 10301002), NSF of Guangdong Province (5005930), and University Special Research Fund for Ph.D. Program (20060574002).

http://www.siam.org/journals/sicon/48-2/70368.html

[†] Department of Math, National University of Singapore (NUS), Singapore (matdm@nus.edu.sg, lipeifan@nus.edu.sg). Min Dai is also an affiliated member of RMI, NUS.

[‡] Department of Applied Math, Tongji University, Shanghai, China (jianglsk@yahoo.com.cn).

[§] Department of Math, South China Normal University, Guangzhou, China (fhyi@scnu.edu.cn).

It is challenging to take the finite horizon case into consideration since the corresponding free boundary (optimal strategy) varies with time. Theoretical analysis on the finite horizon problem became possible only very recently. For example, Liu and Loewenstein (2002) examined the optimal strategy by virtue of a sequence of analytical solutions that converge to the solution of the finite horizon optimal investment problem with transaction costs. Dai and Yi (2009) considered the same problem and derived an equivalent variational inequality by which they completely figured out the optimal strategy. Dai, Xu, and Zhou (2007) extended the idea of Dai and Yi (2009) to the continuous-time mean-variance analysis with transaction costs.

So far, the study of the finite horizon problems has been limited to the no-consumption case. In this paper, we will take into account investment and consumption together with finite horizon and transaction costs, and aim to characterize the optimal strategy. Let us first look at the problem formulation.

1.1. Problem formulation. We consider a continuous-time market in which there are only two investment instruments: a bank account and a stock with price dynamics given, respectively, by

$$\begin{aligned} dP_{0t} &= rP_{0t}dt, \\ dP_{1t} &= P_{1t}\left[\alpha dt + \sigma d\mathcal{B}_t\right]. \end{aligned}$$

Here $r > 0$, $\alpha > r$, and $\sigma > 0$ are constants, and the process $\{\mathcal{B}_t;\ t \in [0, T]\}$ is a standard one-dimensional Brownian motion on a filtered probability space $\big(S, \mathscr{F}, \{\mathscr{F}_t\}_{t\in[0,T]}, P\big)$ with $\mathcal{B}_0 = 0$ almost surely. We assume that the filtration $\{\mathscr{F}_t\}_{t\in[0,T]}$ is generated by the Brownian motion and is right-continuous, and that each $\mathscr{F}_t$ contains all P-null sets of $\mathscr{F}$.

Assume that a CRRA investor holds X_t and Y_t in bank and stock, respectively, expressed in monetary terms. In the presence of transaction costs, the equations describing their evolution are

$$dX_t = (rX_t - C_t)\,dt - (1+\lambda)dL_t + (1-\mu)dM_t, \tag{1.1}$$

$$dY_t = \alpha Y_t dt + \sigma Y_t d\mathcal{B}_t + dL_t - dM_t, \tag{1.2}$$

where $C_t \geq 0$ is the consumption rate, L_t and M_t are right-continuous (with left-hand limits), nonnegative, and nondecreasing $\{\mathscr{F}_t\}_{t\in[0,T]}$-adapted processes with $L_0 = M_0 = 0$, representing cumulative dollar values for the purpose of buying and selling stock, respectively. The constants $\lambda \in [0, \infty)$ and $\mu \in [0, 1)$ appearing in these equations account for proportional transaction costs incurred on purchase and sale of stock, respectively, and $\lambda + \mu > 0$.

Without loss of generality,[1] we always assume $Y_t > 0$. Due to transaction costs, the investor's net wealth in monetary terms at time t is

$$W_t = X_t + (1-\mu)Y_t \quad \text{for } Y_t > 0.$$

Since it is required that the investor's net wealth be positive, the solvency region is defined as

$$\mathscr{S} = \left\{(x, y) \in R^2 : x + (1-\mu)y > 0,\ y > 0\right\}.$$

[1] Given $\alpha > r$, it can be shown that a short position in stock is never optimal. A brief discussion is located in Appendix B. So, we need only consider $Y_t \geq 0$. Thanks to continuity, we can confine ourselves to $Y_t > 0$.

Assume that the investor is given an initial position $(x, y) \in \mathscr{S}$ at time 0. An investment and consumption strategy (L, M, C) is admissible for (x, y) starting from time $s \in [0, T)$ if (X_t, Y_t) given by (1.1)–(1.2) with $X_s = x$ and $Y_s = y$ is in $\mathscr{S}$ for all $t \in [s, T]$. We let $A_s(x, y)$ be the set of admissible investment strategies starting from time s.

The investor's problem is to choose an admissible strategy so as to maximize the expected utility of accumulative consumptions and terminal wealth,

$$\sup_{(L,M,C)\in A_0(x,y)} E_0^{x,y}\left[\int_0^T e^{-\beta s}U\left(C_s\right)ds + e^{-\beta T}U(W_T)\right] \tag{1.3}$$

subject to (1.1)–(1.2). Here $\beta > 0$ is the discounting factor, $E_t^{x,y}$ denotes the conditional expectation at time t given that initial endowment $X_t = x$, $Y_t = y$, and the utility function is taken as[2]

$$U(W) = \frac{W^\gamma}{\gamma} \quad \text{for } 0 < \gamma < 1.$$

Problem (1.3) is a singular stochastic control problem for state processes X_t and Y_t due to controls that are allowed to be discontinuous. Let us define the value function by

$$\varphi(x, y, t) = \sup_{(L,M,C)\in A_t(x,y)} E_t^{x,y}\left[\int_t^T e^{-\beta(s-t)}U\left(C_s\right)ds + e^{-\beta(T-t)}U(W_T)\right], \qquad (x, y) \in \mathscr{S},\ t \in [0, T).$$

It turns out that $\varphi(x, y, t)$ satisfies the following Hamilton–Jacobi–Bellman equation (cf. Shreve and Soner (1994), Lai and Lim (2003), and Fleming and Soner (2006)):

$$\min\left\{-\partial_t\varphi - \mathscr{L}\varphi, -(1-\mu)\partial_x\varphi + \partial_y\varphi, (1+\lambda)\partial_x\varphi - \partial_y\varphi\right\} = 0, \qquad (x, y) \in \mathscr{S},\ t \in [0, T), \tag{1.4}$$

with the terminal condition

$$\varphi(x, y, T) = U\left(x + (1-\mu)y\right), \tag{1.5}$$

where

$$\mathscr{L}\varphi = \frac{1}{2}\sigma^2 y^2 \partial_{yy}\varphi + \alpha y \partial_y\varphi + rx\partial_x\varphi - \beta\varphi + \frac{1-\gamma}{\gamma}\left(\partial_x\varphi\right)^{-\frac{\gamma}{1-\gamma}}.$$

Making use of the homotheticity of the utility function, it follows that for any positive constant ρ,

$$\varphi(\rho x, \rho y, t) = \rho^\gamma \varphi\left(x, y, t\right).$$

This inspires us to make a transformation

$$V\left(\frac{x}{y}, \tau\right) \equiv \varphi\left(\frac{x}{y}, 1, t\right) = \frac{1}{y^\gamma}\varphi(x, y, t) \text{ and } \tau = T - t. \tag{1.6}$$

[2] Problem formulation with log utility ($\gamma = 0$) is presented in Appendix A. The corresponding PDE problem is relatively easy to handle. Note that a CRRA utility also allows $\gamma < 0$, in which case we meet a technical difficulty. See Remark 2.2.

The governing equation for $V(\cdot,\cdot)$ is given by

$$(1.7)\qquad \begin{cases} \min\Big\{\partial_\tau V - \mathcal{L}_1 V - \frac{1-\gamma}{\gamma}\left(\partial_x V\right)^{-\frac{\gamma}{1-\gamma}}, \\ \qquad\qquad -(x+1-\mu)\partial_x V + \gamma V,\ (x+1+\lambda)\partial_x V - \gamma V\Big\} = 0, \\ V(x,0) = \frac{1}{\gamma}(x+1-\mu)^\gamma, \quad -(1-\mu) < x < +\infty,\ 0 < \tau \le T, \end{cases}$$

where

$$\mathcal{L}_1 V \equiv \frac{1}{2}\sigma^2 x^2 \partial_{xx} V + \beta_2 x \partial_x V + \gamma\beta_1 V - \beta V$$

with $\beta_2 = -\left(\alpha - r - (1-\gamma)\sigma^2\right)$, $\beta_1 = \alpha - \frac{1}{2}(1-\gamma)\sigma^2$.

Except for the temporal term and terminal condition, problem (1.7) is analogous to that in Davis and Norman (1990) or Shreve and Soner (1994). It can be shown that the problem has a unique viscosity solution (cf. Davis, Panas, and Zariphopoulou (1993), Shreve and Soner (1994), and Fleming and Soner (2006)).

1.2. Gradient constraints. Similar to Dai and Yi (2009), we make use of the transformation $w(x,\tau) = \frac{1}{\gamma}\ln(\gamma V)$ to reduce problem (1.7) to a parabolic variational inequality with gradient constraints:

Problem A.

$$\begin{cases} \min\Big\{\partial_\tau w - \mathcal{L}_2 w - \frac{1-\gamma}{\gamma}\left(e^w \partial_x w\right)^{-\frac{\gamma}{1-\gamma}},\ \frac{1}{x+1-\mu} - \partial_x w,\ \partial_x w - \frac{1}{x+1+\lambda}\Big\} = 0, \\ w(x,0) = \ln\left(x+1-\mu\right), \quad -(1-\mu) < x < +\infty,\ 0 < \tau \le T. \end{cases}$$

Here

$$\mathcal{L}_2 w = \frac{1}{2}\sigma^2 x^2\left(\partial_{xx} w + \gamma\left(\partial_x w\right)^2\right) + \beta_2 x \partial_x w + \beta_1 - \frac{1}{\gamma}\beta.$$

Problem A gives rise to two free boundaries that correspond to the optimal buying and selling strategies. So, our main purpose is to investigate the behaviors of the free boundaries.[3] In addition, we are interested in the regularity of the solution to Problem A.

PDE problems related to Problem A (variational inequalities with gradient constraints) have been studied by many researchers, including Evans (1979), Wiegner (1981), Ishii and Koike (1983), Hu (1986), Soner and Shreve (1991), and Zhu (1992). It can be shown that the solution to this type of problem belongs to $W_p^1 \cap W_{p,loc}^2$, $1 \le p < \infty$ (in the spatial direction). This regularity turns out to be sharp in the absence of convexity. But the present problem does have the convexity. Indeed, Shreve and Soner (1994) and Dai and Yi (2009) have obtained C^2 smoothness (in the spatial direction) for the infinite horizon case and the no-consumption case, respectively. We will show that it is still true for the present problem. However, the viscosity solution approach adopted by Shreve and Soner (1994) seems unable to deal with the present time-dependent problem. On the other hand, it is intractable to study the properties of free boundaries directly from Problem A. Hence, we will follow Dai and Yi (2009) to adopt an indirect approach.

[3] It can be shown that the optimal consumption rate $C = (\partial_x \varphi)^{1/(1-\gamma)}$ in the no-trading region. So, we need only concentrate on the optimal buying and selling strategies.

1.3. Our approach and novelty. We attempt to reduce Problem A to a standard variational inequality. In what follows let us briefly introduce the idea.

As in Dai and Yi (2009), we set

$$v \equiv \partial_x w = \frac{1}{\gamma}\frac{\partial_x V}{V}.$$

Formally we have

$$\begin{aligned} \frac{\partial}{\partial x}\mathcal{L}_2 w &= \frac{1}{2}\sigma^2 x^2 \partial_{xx} v - \left(\alpha - r - (2-\gamma)\,\sigma^2\right) x\partial_x v - \left(\alpha - r - (1-\gamma)\,\sigma^2\right) v \\ &\qquad + \gamma\sigma^2\left(x^2 v\partial_x v + xv^2\right) \\ &\equiv \mathcal{L}v \end{aligned} \tag{1.8}$$

and

$$\begin{aligned} \frac{\partial}{\partial x}\left(-\frac{1-\gamma}{\gamma}\left(e^w \partial_x w\right)^{-\frac{\gamma}{1-\gamma}}\right) &= \left(e^{\gamma w} v\right)^{-\frac{1}{1-\gamma}}\left(v^2 + \partial_x v\right) \\ &\equiv \mathcal{L}_w v. \end{aligned} \tag{1.9}$$

Then we postulate that v is the solution to the following standard variational inequality, also called the double obstacle problem:

$$\begin{cases} \min\left\{\max\left\{\partial_\tau v - \mathcal{L}v + \mathcal{L}_w v, v - \frac{1}{x+1-\mu}\right\}, v - \frac{1}{x+1+\lambda}\right\} = 0, \\ v(x,0) = \frac{1}{x+1-\mu}, \end{cases}$$

or, equivalently,

$$\begin{cases} \partial_\tau v - \mathcal{L}v + \mathcal{L}_w v = 0 & \text{if } \frac{1}{x+1+\lambda} < v < \frac{1}{x+1-\mu}, \\ \partial_\tau v - \mathcal{L}v + \mathcal{L}_w v \le 0 & \text{if } v = \frac{1}{x+1-\mu}, \\ \partial_\tau v - \mathcal{L}v + \mathcal{L}_w v \ge 0 & \text{if } v = \frac{1}{x+1+\lambda}, \\ v(x,0) = \frac{1}{x+1-\mu} \end{cases} \tag{1.10}$$

in $-(1-\mu) < x < +\infty$, $0 < \tau \le T$. Here $\frac{1}{x+1+\lambda}$ and $\frac{1}{x+1-\mu}$ stand for lower and upper obstacles, respectively. It is worth pointing out that $\mathcal{L}v$ is degenerate on $x = 0$. Compared with Dai and Yi (2009), (1.10) has an additional term $\mathcal{L}_w v$ which will cause much technical difficulty.

It is well known that the solution to a double obstacle problem is of C^1 in the spatial direction. We immediately obtain $w \in C^2$ in the spatial direction (except on $x = 0$) provided that $v = \partial_x w$ satisfies (1.10). More importantly, we will see later that it is rather straightforward to analyze the behaviors of free boundaries in terms of problem (1.10).

Hence, the main task is to prove the equivalence between Problem A and problem (1.10), which essentially indicates the connection between a singular control problem and an optimal stopping problem (cf. Karatzas and Shreve (1984) and Soner and Shreve (1991)). In the no-consumption case in which the counterpart of (1.10) does not contain the nonlinear term $\mathcal{L}_w v$, Dai and Yi (2009) first established such an equivalence in terms of which they completely characterized the optimal buying and selling strategies. Nevertheless, the equivalence has never been revealed for the present problem with consumption. We would like to emphasize that it is not an easy task to establish the equivalence. One of the main barriers is that $\mathcal{L}_w v$ depends on w, which leads problem (1.10) to not be a self-contained system. We will exploit an auxiliary condition with which the problem (1.10) can be shown to have a solution by the Schauder fixed point theorem and to be equivalent to Problem A. In addition, it can be proved in the no-consumption case that $\mathcal{L}v$ is hypoelliptic, which enables us to take into account (1.10) independently in $x > 0$ and $x < 0$. However, it is no longer true when a consumption is taken into consideration. Hence, we will have to introduce a regularized problem in order to deal with the degeneracy of $\mathcal{L}v$ on $x = 0$.

The rest of the paper is arranged as follows. In the next section, we take into account the regularity of the solution to problem (1.10) with a known $w(x,t)$. In section 3, we derive the auxiliary condition with which problem (1.10) becomes self-contained and has a solution by use of the regularity result obtained in section 2 and the Schauder fixed point theorem. In section 4, we make use of problem (1.10) to investigate the behaviors of the free boundaries. Section 5 is devoted to the equivalence between Problem A and problem (1.10) with the auxiliary condition. We conclude in section 6.

2. The problem (1.10) with a known $w(x,\tau)$.

In this section, we study problem (1.10) with known $w(x,\tau)$ which is assumed to possess the following properties:

$$|w(x,\tau) - \ln(x+1-\mu)| \le M_T, \tag{2.1}$$

$$\frac{1}{x+1+\lambda} \le \partial_x w(x,\tau) \le \frac{1}{x+1-\mu}, \tag{2.2}$$

$$|\partial_\tau w(x,\tau)| \le M, \tag{2.3}$$

$$w(x,0) = \ln(x+1-\mu). \tag{2.4}$$

Here M and M_T are positive constants.

Notice that the initial value and the upper obstacle in (1.10) are unbounded near $x = -(1-\mu)$. As a result, we confine problem (1.10) to the domain $\Omega_T = (x^*, +\infty) \times (0,T)$ with a boundary condition

$$\partial_x v(x^*,\tau) = -\frac{1}{(x^*+1-\mu)^2}, \qquad \tau \in (0,T), \tag{2.5}$$

where $x^* \in (-(1-\mu), 0)$. We always assume x^* to be close enough to $-(1-\mu)$. Later we will see that it is without loss of generality.

In addition, owing to the unboundedness of Ω_T, we further confine problem (1.10) to a bounded domain $\Omega_T^R = (x^*, R) \times (0,T)$ with $R > 0$. On $x = R$ we impose a boundary condition

$$\partial_x v(R,\tau) + v^2(R,\tau) = 0, \qquad \tau \in (0,T). \tag{2.6}$$

Since the operator $\mathcal{L}$ is degenerate on $x = 0$, we instead consider the following regularized problem: for $\delta > 0$,

$$\begin{cases} \partial_\tau v_\delta - \mathcal{L}_\delta v_\delta + \mathcal{L}_w v_\delta = 0 & \text{if } \frac{1}{x+1+\lambda} < v_\delta < \frac{1}{x+1-\mu}, \\ \partial_\tau v_\delta - \mathcal{L}_\delta v_\delta + \mathcal{L}_w v_\delta \ge 0 & \text{if } v_\delta = \frac{1}{x+1+\lambda}, \\ \partial_\tau v_\delta - \mathcal{L}_\delta v_\delta + \mathcal{L}_w v_\delta \le 0 & \text{if } v_\delta = \frac{1}{x+1-\mu}, \\ \partial_x v_\delta(x^*,\tau) = -\frac{1}{(x^*+1-\mu)^2}, & \\ \partial_x v_\delta(R,\tau) + v_\delta^2(R,\tau) = 0, & \qquad (x,\tau) \in \Omega_T^R, \\ v_\delta(x,0) = \frac{1}{x+1-\mu}, & \end{cases} \tag{2.7}$$

where

$$\mathcal{L}_\delta v_\delta = \mathcal{L} v_\delta + \delta \partial_{xx} v_\delta.$$

LEMMA 2.1. *For a given $w(x,t)$ satisfying* (2.1)–(2.4)*, problem* (2.7) *has a solution $v_\delta \in W_p^{2,1}(\Omega_T^R)$, $1 < p < +\infty$, and*

$$\frac{1}{x+1+\lambda} \le v_\delta \le \frac{1}{x+1-\mu}, \tag{2.8}$$

$$-\frac{K}{(x+1-\mu)^2} \le \partial_x v_\delta \le -v_\delta^2, \tag{2.9}$$

where K is a positive constant independent of δ and R.

Proof. By making use of the standard penalty method and the fixed point theorem (cf. Friedman (1982), section 1.8), we can show that problem (2.7) has a solution in $W_p^{2,1}(\Omega_T^R)$, $1 < p < +\infty$. Inequality (2.8) is apparent. We now prove (2.9). Let us first consider the right-hand side inequality. Clearly

$$\partial_x v_\delta + v_\delta^2 = 0 \qquad \text{if } \ v_\delta = \frac{1}{x+1+\lambda} \ \text{ or } \ v_\delta = \frac{1}{x+1+\mu}.$$

So, we need only show $\partial_x v_\delta + v_\delta^2 \le 0$ in $\mathcal{M}$, where

$$\mathcal{M} = \left\{ (x,\tau) \in \Omega_T^R : \frac{1}{x+1+\lambda} < v_\delta < \frac{1}{x+1-\mu} \right\}.$$

Denote $p(x,\tau) = \partial_x v_\delta(x,\tau)$ and $q(x,\tau) = v_\delta^2(x,\tau)$, then

$$\begin{aligned}
&\partial_\tau p - \left(\frac{1}{2}\sigma^2 x^2 + \delta\right) \partial_{xx} p + \left(\alpha - r - (3-\gamma)\sigma^2\right) x \partial_x p + \left(2\alpha - 2r - (3-2\gamma)\sigma^2\right) p \\
&\quad + (e^{\gamma w} v_\delta)^{-\frac{1}{1-\gamma}} (\partial_x q + \partial_x p) + \partial_x[(e^{\gamma w} v_\delta)^{-\frac{1}{1-\gamma}}](q+p) \\
&\quad = \gamma\sigma^2 \left(4xv\partial_x v_\delta + x^2 (\partial_x v_\delta)^2 + x^2 v \partial_{xx} v_\delta + v_\delta^2\right) \qquad \text{in } \ \mathcal{M}
\end{aligned} \tag{2.10}$$

and

$$\begin{aligned}
&\partial_\tau q - \left(\frac{1}{2}\sigma^2 x^2 + \delta\right) \partial_{xx} q + \left(\alpha - r - (2-\gamma)\sigma^2\right) x \partial_x q + \left(2\alpha - 2r - (2-2\gamma)\sigma^2\right) q \\
&\quad + 2v_\delta (e^{\gamma w} v_\delta)^{-\frac{1}{1-\gamma}} (q+p) \\
&\quad = -\sigma^2 x^2 (\partial_x v_\delta)^2 + \gamma\sigma^2 \left(2x^2 v_\delta^2 \partial_x v_\delta + 2x v_\delta^3\right) - \delta\sigma^2 (\partial_x v_\delta)^2 \qquad \text{in } \ \mathcal{M}.
\end{aligned}$$

Let $H(x,\tau) = p(x,\tau) + q(x,\tau)$. It is not hard to verify

$$\begin{aligned}
&\partial_\tau H - \left(\frac{1}{2}\sigma^2 x^2 + \delta\right) \partial_{xx} H + \left(\alpha - r - (3-\gamma)\sigma^2 - \gamma\sigma^2 x v_\delta\right) x \partial_x H \\
&\quad + \left(2\alpha - 2r - (3-2\gamma)\sigma^2 - 2\gamma\sigma^2 x v_\delta\right) H \\
&\quad + (e^{\gamma w} v_\delta)^{-\frac{1}{1-\gamma}} \partial_x H + \{\partial_x[(e^{\gamma w} v_\delta)^{-\frac{1}{1-\gamma}}] + 2v_\delta (e^{\gamma w} v_\delta)^{-\frac{1}{1-\gamma}}\} H \\
&\quad = -(1-\gamma)\sigma^2 (x\partial_x v_\delta + v_\delta)^2 - \delta\sigma^2 (\partial_x v_\delta)^2 \le 0 \qquad \text{in } \ \mathcal{M}.
\end{aligned}$$

Apparently $H \leq 0$ on $\partial\mathcal{M} \cap (\{x = x^*\} \cup \{x = R\} \cup \{\tau = 0\})$. In terms of the maximum principle (cf. Friedman (1982), p. 74), we then deduce $H \leq 0$ in $\mathcal{M}$.

Now we turn to the proof of the left-hand side inequality of (2.9). Note that (2.10) can be rewritten as

$$\partial_\tau p - \mathcal{T}p = 0 \text{ in } \mathcal{M},$$

where

$$\begin{aligned}\mathcal{T}p = {} & \left(\frac{1}{2}\sigma^2 x^2 + \delta\right)\partial_{xx}p - (\alpha - r - (3-\gamma)\sigma^2)x\partial_x p - (2\alpha - 2r - (3-2\gamma)\sigma^2)p \\ & + \gamma\sigma^2(x^2 v_\delta \partial_x p + x^2 p^2 + 4x v_\delta p + v_\delta^2) \\ & + \frac{1}{1-\gamma}(e^{\gamma w} v_\delta)^{-\frac{1}{1-\gamma}-1} e^{\gamma w}(\gamma \partial_x w v_\delta + p)(v_\delta^2 + p) \\ & - (e^{\gamma w} v_\delta)^{-\frac{1}{1-\gamma}}(2 v_\delta p + \partial_x p).\end{aligned}$$

It can be verified that for constant K big enough,

$$\begin{aligned}& (\partial_\tau - \mathcal{T})\left(-\frac{K}{(x+1-\mu)^2}\right) \\ = {} & -\frac{1}{(x+1-\mu)^4}\left[\gamma\sigma^2 x^2 + \frac{1}{1-\gamma}(e^{\gamma w} v_\delta)^{-\frac{1}{1-\gamma}-1} e^{\gamma w}\right]K^2 \\ & + \frac{1}{(x+1-\mu)^2}\left\{\frac{3\sigma^2 x^2 + 6\delta}{(x+1-\mu)^2} + 2\frac{(\alpha - r - (3-\gamma)\sigma^2)x - \gamma\sigma^2 x^2 v_\delta}{x+1-\mu}\right. \\ & \qquad - \left(2\alpha - 2r - (3-2\gamma)\sigma^2\right) + 4\gamma\sigma^2 x v_\delta \\ & \qquad \left. + (e^{\gamma w} v_\delta)^{-\frac{1}{1-\gamma}}\left[\frac{1}{1-\gamma}\left(\gamma \partial_x w + v_\delta\right) + 2\left(\frac{1}{(x+1-\mu)} - v_\delta\right)\right]\right\}K \\ & - \frac{\gamma}{1-\gamma}(e^{\gamma w} v_\delta)^{-\frac{1}{1-\gamma}}\partial_x w v_\delta^2 - \gamma\sigma^2 v_\delta^2 \\ \leq {} & 0,\end{aligned}$$

because the coefficient of the leading term K^2 is negative. Here K is independent of δ and R. It is clear that $p \geq -\frac{K}{(x+1-\mu)^2}$ on $\partial\mathcal{M} \cap (\{x = x^*\} \cup \{x = R\} \cup \{\tau = 0\})$. Again, applying the maximum principle yields the desired result. Therefore, the proof is complete. □

Remark 2.2. We point out that the above proof for the left-hand side inequality of (2.9) requires $\gamma > 0$. All results in this paper could be extended to the case of $\gamma < 0$ if the condition is removed.

LEMMA 2.3. *For a given* $w(x,t)$ *satisfying* (2.1)–(2.4), *problem* (1.10) *confined to* $\overline{\Omega}_T^R$ *with boundary conditions* (2.5)–(2.6) *has a solution* $v \in W_p^{2,1}(\Omega_T^R \backslash \{-\eta < x < \eta\}) \cap C(\overline{\Omega}_T^R)$ *for any small* $\eta > 0$, $1 < p < +\infty$, *and*

$$\frac{1}{x+1+\lambda} \leq v \leq \frac{1}{x+1-\mu}, \tag{2.11}$$

$$-\frac{K}{(x+1-\mu)^2} \leq \partial_x v \leq -v^2, \tag{2.12}$$

$$|v(x,\cdot)|_{C^{\theta/2}[0,T]} \leq C, \tag{2.13}$$

where K, θ, and C are positive constants independent of R, $0 < \theta < 1$, and $C^{\theta/2}[0,T]$ is a Hölder space.

Proof. Let v be the limit of a weakly convergent subsequence of $\{v_\delta\}$ as $\delta \to 0$. We immediately get (2.11)–(2.12). Now we prove (2.13). When $x > 0$, letting $\delta \to 0$ in (2.7), we infer that (1.10) holds in $(0,R) \times (0,T)$ and can be rewritten as

$$\begin{cases} \partial_\tau v - \mathcal{L}v = f(x,\tau), & (x,\tau) \in (0,R) \times (0,T), \\ \partial_x v(R,\tau) + v^2(R,\tau) = 0, & \\ v(x,0) = \frac{1}{x+1-\mu}, & \end{cases} \tag{2.14}$$

where

$$\begin{aligned} f(x,\tau) = &-(e^{\gamma w}v)^{-\frac{1}{1-\gamma}}(v^2 + \partial_x v) \\ &+ \chi_{\{v=\frac{1}{x+1-\mu}\}} \frac{1}{(x+1-\mu)^3}[(\alpha-r)x + (1-\mu)(\alpha - r - (1-\gamma)\sigma^2)] \\ &+ \chi_{\{v=\frac{1}{x+1+\lambda}\}} \frac{1}{(x+1+\lambda)^3}[(\alpha-r)x + (1+\lambda)(\alpha - r - (1-\gamma)\sigma^2)] \end{aligned}$$

and χ_A is the indicator function on set A. Notice that $f(x,\tau)$ has a bound independent of R.

By transformation

$$x = e^y, \quad v(x,\tau) = u(y,\tau),$$

problem (2.14) becomes

$$\begin{cases} \partial_\tau u - \mathcal{L}_y u = g(y,\tau), & (y,\tau) \in (-\infty, \ln R) \times (0,T), \\ \partial_y u(\ln R,\tau) = -Ru^2(\ln R,\tau), & \\ u(y,0) = \frac{1}{e^y+1-\mu}, & \end{cases} \tag{2.15}$$

where

$$\mathcal{L}_y u = \frac{1}{2}\sigma^2 \partial_{yy}u - \left(\alpha - r - \left(\frac{3}{2} - \gamma\right)\sigma^2\right)\partial_y u - (\alpha - r - (1-\gamma)\sigma^2)u + \gamma\sigma^2 (e^y u)(\partial_y u + u)$$

and $g(y,\tau) = f(e^y,\tau)$ is still a bounded function.

Owing to (2.11), we infer that both the boundary condition and the coefficients of $\mathcal{L}_y u$ are bounded. By applying the $C^{\theta,\theta/2}$ $(0 < \theta < 1)$ estimate of the parabolic equation, we obtain

$$|u|_{C^{\theta,\theta/2}((-\infty,\ln R]\times[0,T])} \le C,$$

where C is independent of R. Especially $|u(y,\cdot)|_{C^{\theta/2}[0,T]} \le C$ or

$$|v(x,\cdot)|_{C^{\theta/2}[0,T]} \le C, \qquad 0 \le x \le R.$$

In the same way,

$$|v(x,\cdot)|_{C^{\theta/2}[0,T]} \le C, \qquad x^* \le x \le 0.$$

Thanks to (2.12), v is continuous with respect to x. We then obtain (2.13) and $v \in C(\overline{\Omega}_T^R)$.

It remains to show that v is the solution to (1.10) in Ω_T^R with boundary conditions (2.5)–(2.6). In fact, we need only prove that (1.10) holds near $x = 0$ in the distributional sense. For any $(0, \tau_0)$, let us first consider the case

$$\frac{1}{1+\lambda} < v\left(0, \tau_0\right) < \frac{1}{1-\mu}.$$

Due to the continuity of v, there exist $\varepsilon > 0$ and $x_1 < 0 < x_2$ such that

$$\frac{1}{x_2+1+\lambda} < v\left(x_2, \tau\right) < v\left(x_1, \tau\right) < \frac{1}{x_1+1-\mu} \quad \text{for } |\tau - \tau_0| < \varepsilon.$$

For fixed x_1, $v_\delta\left(x_1, \tau\right)$ uniformly converges to $v\left(x_1, \tau\right)$ for $|\tau - \tau_0| < \varepsilon$. So, there is $\delta_0 > 0$ such that

$$v_\delta\left(x_1, \tau_1\right) < \frac{1}{x_1+1-\mu} \quad \text{for } |\tau - \tau_0| < \varepsilon,\ \delta < \delta_0. \tag{2.16}$$

In the same way, for fixed $x_2 > 0$,

$$v_\delta\left(x_2, \tau_2\right) > \frac{1}{x_2+1+\lambda} \quad \text{for } |\tau - \tau_0| < \varepsilon,\ \delta < \delta_0. \tag{2.17}$$

Note that (2.16) can be rewritten as

$$\left(x_1+1-\mu\right)^2 v_\delta(x_1, \tau) < x_1+1-\mu \quad \text{for } |\tau - \tau_0| < \varepsilon,\ \delta < \delta_0,$$

and

$$\begin{aligned} &\partial_x \left((x+1-\mu)^2 v_\delta - (x+1-\mu) \right) \\ &\quad = -\left[(x+1-\mu) v_\delta - 1 \right]^2 + (x+1-\mu)^2 \left(\partial_x v_\delta + v_\delta^2 \right) \le 0, \end{aligned} \tag{2.18}$$

where we have used the right-hand side inequality in (2.9). We then deduce

$$(x+1-\mu)^2 v_\delta(x, \tau) < x+1-\mu \quad \text{for } |\tau - \tau_0| < \varepsilon,\ x_1 < x < x_2,\ \delta < \delta_0,$$

namely,

$$v_\delta(x, \tau) < \frac{1}{x+1-\mu} \quad \text{for } |\tau - \tau_0| < \varepsilon,\ x_1 < x < x_2,\ \delta < \delta_0. \tag{2.19}$$

On the other hand, due to

$$\partial_x \left(v_\delta - \frac{1}{x+1+\lambda} \right) = \partial_x v_\delta + \frac{1}{(x+1+\lambda)^2} \le \partial_x v_\delta + v_\delta^2 \le 0, \tag{2.20}$$

it follows from (2.17) that

$$v_\delta\left(x, \tau\right) > \frac{1}{x+1+\lambda} \quad \text{for } |\tau - \tau_0| < \varepsilon,\ x_1 < x < x_2,\ \delta < \delta_0. \tag{2.21}$$

From (2.19) and (2.21), we infer that the first equation of (2.7) holds in $E \equiv \{x_1 < x < x_2,\ |\tau - \tau_0| < \varepsilon\}$. We then deduce by letting $\delta \to 0$ that the first equation of (1.10) holds in E in the distributional sense. Now let us move on to the case

$v\left(0,\tau_0\right)=\frac{1}{1+\lambda}$. Note that $v\left(0,\tau_0\right)<\frac{1}{1-\mu}$. Using a similar argument, we deduce that there is a neighborhood E of $(0,\tau_0)$ such that

$$v_\delta\left(x,\tau\right)<\frac{1}{x+1-\mu},\ \ (x,\tau)\in E$$

when δ is sufficiently small. Then,

$$\partial_\tau v_\delta-\mathcal{L}_\delta v_\delta+\mathcal{L}_w v_\delta\geq 0 \text{ in } E.$$

Again, we let $\delta\to 0$ to get the desired result. The case of $v\left(0,\tau_0\right)=\frac{1}{1-\mu}$ is similar. This completes the proof. □

Thanks to the right-hand side inequality in (2.12), we infer that (2.18) and (2.20) are valid for v. This indicates that there are two functions $x_{s,w}(\tau)$ and $x_{b,w}(\tau)$, $\tau\in(0,T]$, such that

$$\left\{(x,\tau)\in\Omega_T^R: v=\frac{1}{x+1-\mu}\right\}=\left\{(x,t)\in\Omega_T^R: x\leq x_{s,w}(\tau)\right\}, \tag{2.22}$$

$$\left\{(x,\tau)\in\Omega_T^R: v=\frac{1}{x+1+\lambda}\right\}=\left\{(x,t)\in\Omega_T^R: x\geq x_{b,w}(\tau)\right\}. \tag{2.23}$$

For later use, we introduce a lemma.

Lemma 2.4. *Denote* $x_M=-\frac{\alpha-r-(1-\gamma)\sigma^2}{\alpha-r}$ *and assume that* $x^*\in(-(1-\mu),(1-\mu)x_M)$*. Then*

$$x_{s,w}(\tau)\leq\ x_{s,w}(0)\equiv\lim_{\tau\to 0^+}x_{s,w}\left(\tau\right)=(1-\mu)x_M. \tag{2.24}$$

Moreover, $x_{s,w}(\tau)\in C^\infty$ *when* $x_{s,w}(\tau)>x^*$.

Proof. Let us prove (2.24) first. Note that for any $x<x_{s,w}(\tau)$,

$$\begin{aligned}0&\geq\left(\frac{\partial}{\partial\tau}-\mathcal{L}+\mathcal{L}_w\right)\left(\frac{1}{x+1-\mu}\right)\\&=\frac{1-\mu}{\left(x+1-\mu\right)^3}\left[\left(\alpha-r\right)x+\left(1-\mu\right)\left(\alpha-r-\left(1-\gamma\right)\sigma^2\right)\right],\end{aligned}$$

from which we infer $x_{s,w}\left(\tau\right)\leq(1-\mu)x_M$. To show $x_{s,w}(0)=(1-\mu)x_M$, we use the method of contradiction. Suppose not; we would have $x_{s,w}\left(0\right)<(1-\mu)x_M$. Then for any $x_0\in(x_{s,w}(0),(1-\mu)x_M)$, applying the equation $\partial_\tau v-\mathcal{L}v+\mathcal{L}_w v=0$ at $\tau=0$ gives $\partial_\tau v|_{\tau=0,\,x=x_0}=\mathcal{L}v-\mathcal{L}_w v|_{\tau=0,\,x=x_0}=\mathcal{L}\left(\frac{1}{x+1-\mu}\right)\big|_{x=x_0}>0$, which conflicts with the apparent fact $\partial_\tau v|_{\tau=0}\leq 0$.

Using (2.12) and analogous arguments as in Dai, Xu, and Zhou (2007), we can show $x_{s,w}(\tau)\in C^\infty$ when $x_{s,w}(\tau)>x^*$. □

3. The problem (1.10) with an auxiliary condition. As mentioned before, we need an auxiliary condition to make problem (1.10) self-contained. Now let us exploit the condition. Assume that $v=\partial_x w$ is a solution to problem (1.10) in Ω_T. Due to Lemma 2.4, we expect that there would be a function $x_s\left(\tau\right):(0,T)\to(x^*,+\infty)$, such that

$$\left\{(x,\tau)\in\Omega_T: v=\frac{1}{x+1-\mu}\right\}=\left\{(x,\tau)\in\Omega_T: x\leq x_s\left(\tau\right)\right\}.$$

So, we have $w(x,\tau) = A(\tau) + \ln(x + 1 - \mu)$, $x \le x_s(\tau)$, where $A(0) = 0$ and $A(\tau)$, $\tau > 0$, is to be determined. Then we conjecture

$$\begin{aligned} w(x,\tau) &= w(x_s(\tau),\tau) + \int_{x_s(\tau)}^{x} v(\xi,\tau)d\xi \\ &= A(\tau) + \ln(x_s(\tau) + 1 - \mu) + \int_{x_s(\tau)}^{x} v(\xi,\tau)\,d\xi \quad \text{for any} \quad (x,\tau) \in \Omega_T. \end{aligned}$$

It is expected that $v(\cdot,\tau) \in C^1$ and $w(\cdot,\tau) \in C^2$. Thus, we should have

$$\partial_x w|_{x=x_s(\tau)} = \frac{1}{x_s(\tau) + 1 - \mu}, \ \partial_{xx} w|_{x=x_s(\tau)} = -\frac{1}{(x_s(\tau) + 1 - \mu)^2},$$

which yields

$$\begin{aligned} A'(\tau) = \partial_\tau w(x_s(\tau),\tau) &= \mathcal{L}_2 w + \frac{1-\gamma}{\gamma}\left(e^w \partial_x w\right)^{-\frac{\gamma}{1-\gamma}}\Big|_{x=x_s(\tau)} \\ &= \frac{1-\gamma}{\gamma} e^{-\frac{\gamma}{1-\gamma}A(\tau)} + f(x_s(\tau)). \end{aligned} \tag{3.1}$$

Here

$$f(x) = \frac{1}{(x+1-\mu)^2}\left[rx^2 + (\alpha + r)(1-\mu)x + \left(\alpha - \frac{1}{2}(1-\gamma)\sigma^2\right)(1-\mu)^2\right] - \frac{1}{\gamma}\beta.$$

Notice that (3.1) can be rewritten as

$$\left(e^{\frac{\gamma}{1-\gamma}A(\tau)}\right)' = \frac{\gamma}{1-\gamma} f(x_s(\tau)) e^{\frac{\gamma}{1-\gamma}A(\tau)} + 1.$$

Combining with $A(0) = 0$, we obtain

$$\begin{aligned} A(\tau) &= \frac{1-\gamma}{\gamma}\log\left[e^{\frac{\gamma}{1-\gamma}\int_0^\tau f(x_s(\zeta))d\zeta} + \int_0^\tau e^{\frac{\gamma}{1-\gamma}\int_{\overline{\tau}}^\tau f(x_s(\zeta))d\zeta} d\overline{\tau}\right] \\ &\equiv \mathcal{H}(x_s(\tau)). \end{aligned}$$

This is the auxiliary condition with which we want to combine the problem (1.10). In other words, we plan to study the following problem.

Problem B. *Find* $w(x,\tau)$, $v(x,\tau)$, *and* $x_s(\tau) : (0,T) \to (x^*, +\infty)$ *such that*

(i) $\left\{(x,\tau) \in \Omega_T : v(x,\tau) = \frac{1}{x+1-\mu}\right\} = \{(x,\tau) \in \Omega_T : x \le x_s(\tau)\}$;

(ii) $v(x,\tau)$, $(x,\tau) \in \Omega_T$ *satisfies* (1.10) *in which*

$$w(x,\tau) = A(\tau) + \ln(x_s(\tau) + 1 - \mu) + \int_{x_s(\tau)}^{x} v(\xi,\tau)\,d\xi, \tag{3.2}$$

where $A(\tau) = \mathcal{H}(x_s(\tau))$.

Proposition 3.1. *Problem* B *allows a unique solution* $(w(x,\tau), v(x,\tau), x_s(\tau))$ *satisfying* (2.1)–(2.4), (2.11)–(2.13), *and* (2.24), *respectively.*

Proof. The uniqueness of the solution is apparent. In the following we will prove the existence of the solution by virtue of the Schauder fixed point theorem. To begin, let us still confine ourselves to a bounded domain $\overline{\Omega}_T^R$. Consider a Banach space $\mathcal{B} = C(\overline{\Omega}_T^R)$ and define

$$\mathcal{D} = \left\{ w(x,\tau) \in \mathcal{B} \,\middle|\, |w(x,\tau) - \ln(x+1-\mu)| \le M_T, \right.$$
$$\frac{1}{x+1+\lambda} \le \partial_x w(x,\tau) \le \frac{1}{x+1-\mu},$$
$$\left. |\partial_\tau w(x,\tau)| \le M, \; w(x,0) = \ln(x+1-\mu) \right\},$$

where M and M_T are positive constants to be prescribed, and $\partial_x w$ and $\partial_\tau w$ are weak derivatives. Clearly $\mathcal{D}$ is a compact convex set in $\mathcal{B}$.

For any $w(x,\tau) \in \mathcal{D}$ given, let $v(x,\tau)$ be the solution of problem (1.10) confined to $\overline{\Omega}_T^R$ with boundary conditions (2.5)–(2.6), and let $x_{s,w}(\tau)$ be the corresponding free boundary as given in (2.22). Define a mapping $\mathcal{F} : \mathcal{D} \to \mathcal{B}$ as follows:

$$\mathcal{F}w = \overline{w}(x,\tau) \equiv A(\tau) + \ln(x_{s,w}(\tau) + 1 - \mu) + \int_{x_{s,w}(\tau)}^{x} v(\xi,\tau)d\xi, \tag{3.3}$$

where $A(\tau) = \mathcal{H}\left(x_{s,w}(\tau)\right)$.

In the following we shall prove $\overline{w}(x,\tau) \in \mathcal{D}$. By definition, it is obvious that $\overline{w}(x,0) = \ln(x+1-\mu)$, $\partial_x \overline{w}(x,\tau) = v(x,\tau)$, and thus $\frac{1}{x+1+\lambda} \le \partial_x \overline{w}(x,\tau) \le \frac{1}{x+1-\mu}$. Using (3.3) and $\frac{1}{x+1+\lambda} \le \partial_x v \le \frac{1}{x+1-\mu}$, it is easily seen that

$$A(\tau) + \ln\frac{x+1+\lambda}{x+1-\mu} + \ln\frac{x_{s,w}(\tau)+1-\mu}{x_{s,w}(\tau)+1+\lambda} \le \overline{w}(x,\tau) - \ln\left(x+1-\mu\right) \le A(\tau). \tag{3.4}$$

According to the definition of $A(\tau)$ and (2.24), $A(\tau)$ is bounded. Then we deduce that there is a positive constant, denoted by M_T independent of R, such that

$$|\overline{w}(x,\tau) - \ln\left(x+1-\mu\right)| < M_T.$$

It remains to show that

$$|\partial_\tau \overline{w}(x,\tau)| \le M. \tag{3.5}$$

By (3.3),

$$\begin{aligned}
&\partial_\tau \overline{w}(x,\tau)\\
&= A'(\tau) + \int_{x_{s,w}(\tau)}^{x} \partial_\tau v(\xi,\tau)d\xi = A'(\tau) + \int_{x_{s,w}(\tau)}^{x} \mathcal{L}v(\xi,\tau)d\xi - \int_{x_{s,w}(\tau)}^{x} \mathcal{L}_w v(\xi,\tau)d\xi\\
&= A'(\tau) + \int_{x_{s,w}(\tau)}^{x} \frac{\partial}{\partial\xi}\mathcal{L}_2\overline{w}(\xi,\tau)d\xi - \int_{x_{s,w}(\tau)}^{x} \mathcal{L}_w v(\xi,\tau)d\xi\\
&= \frac{1-\gamma}{\gamma}e^{-\frac{\gamma}{1-\gamma}A(\tau)} + \frac{1}{2}\sigma^2 x^2(\partial_x v + \gamma v^2) + \beta_2 x v + \beta_1 - \frac{1}{\gamma}\beta - \int_{x_{s,w}(\tau)}^{x} \mathcal{L}_w v(\xi,\tau)d\xi.
\end{aligned} \tag{3.6}$$

Combining with (2.11)–(2.12) and the boundedness of $A(\tau)$, we deduce that there is a constant $M_1 > 0$ independent of R such that

$$\left| \frac{1-\gamma}{\gamma}e^{-\frac{\gamma}{1-\gamma}A(\tau)} + \frac{1}{2}\sigma^2 x^2(\partial_x v + \gamma v^2) + \beta_2 x v + \beta_1 - \frac{1}{\gamma}\beta \right| \le M_1.$$

Regarding the boundedness of the last term in (3.6), observe that $w(x,t)$ has a bound depending only on R. Hence, it is easy to see that there is a constant M_2 depending only on R such that

$$0 \le -\int_{x_{s,w}(\tau)}^{x} \mathcal{L}_w v(\xi,\tau) d\xi \le M_2.$$

We then choose $M = M_1 + M_2$ to obtain (3.5).

So far we have obtained $\mathcal{F}(\mathcal{D}) \subset \mathcal{D}$. Owing to the uniqueness of solution, $\mathcal{F}$ must be a one-one mapping. Thanks to the compactness of $\mathcal{D}$, we then infer that $\mathcal{F}$ must be continuous. Applying the Schauder fixed point theorem we see that Problem B confined to $\overline{\Omega}_T^R$ allows a solution (w_R, v_R, x_s).

To extend the result to domain $\overline{\Omega}_T$, we need only show that $\partial_\tau w_R$ has a uniform bound (i.e., independent of R). Thanks to (1.9) and $\partial_x w_R = v_R$,

$$-\int_{x_s(\tau)}^{x} \mathcal{L}_{w_R} v_R(\xi,\tau) d\xi = \int_{x_s(\tau)}^{x} \frac{\partial}{\partial \xi}\left(\frac{1-\gamma}{\gamma}\left(e^{w_R} v_R\right)^{-\frac{\gamma}{1-\gamma}}\right) d\xi.$$

Combining with (3.6), we obtain

$$\partial_\tau w_R(x,\tau) = \frac{1}{2}\sigma^2 x^2 (\partial_x v + \gamma v^2) + \beta_2 x v + \beta_1 - \frac{1}{\gamma}\beta + \frac{1-\gamma}{\gamma}\left(e^{w_R} v_R\right)^{-\frac{\gamma}{1-\gamma}}.$$

As a result, it suffices to show that $e^{w_R} v_R$ has a uniform bound. Similar to (3.4), we have

$$A(\tau) + \log(x+1+\lambda) + \log\frac{x_s(\tau)+1-\mu}{x_s(\tau)+1+\lambda} \le w_R(x,\tau) \le A(\tau) + \log(x+1-\mu).$$

Owing to $\frac{1}{x+1+\lambda} \le v_R(x,\tau) \le \frac{1}{x+1-\mu}$, we then infer

$$\frac{x_s(\tau)+1-\mu}{x_s(\tau)+1+\lambda} e^{A(\tau)} \le e^{w_R} v_R \le e^{A(\tau)},$$

which is desired. The proof is complete. □

In contrast to $x_{b,w}(\tau)$ in (2.23), we can similarly define the boundary $x_b(\tau)$ related to Problem B as follows:

$$\{(x,t) \in \Omega_T : x \ge x_b(\tau)\} = \left\{(x,\tau) \in \Omega_T : v = \frac{1}{x+1+\lambda}\right\}.$$

4. Behaviors of free boundaries. The equivalence proof between Problems A and B is deferred to section 5. In this section we study the behaviors of free boundaries $x_s(\tau)$ and $x_b(\tau)$ which reflect the optimal selling and buying boundaries, respectively. For comparison, let us first recall the results when there is no consumption.

4.1. Without consumption. In the absence of consumption, the counterpart of (1.10) becomes (see Dai and Yi (2009))

$$(4.1) \qquad \begin{cases} \partial_\tau \overline{v} - \mathcal{L}\overline{v} = 0 & \text{if } \frac{1}{x+1+\lambda} < \overline{v} < \frac{1}{x+1-\mu}, \\ \partial_\tau v - \mathcal{L}\overline{v} \le 0 & \text{if } \overline{v} = \frac{1}{x+1-\mu}, \\ \partial_\tau \overline{v} - \mathcal{L}\overline{v} \ge 0 & \text{if } \overline{v} = \frac{1}{x+1+\lambda}, \\ \overline{v}(x,0) = \frac{1}{x+1-\mu}, & \qquad (x,\tau) \in \Omega_T. \end{cases}$$

In contrast to (1.10), the nonlinear operator $\mathcal{L}_w$ disappears. Problem (4.1) also allows two free boundaries, denoted by $\overline{x}_s(\tau)$ and $\overline{x}_b(\tau)$, such that

$$\left\{(x,\tau)\in\Omega_T:\overline{v}(x,\tau)=\frac{1}{x+1-\mu}\right\}=\left\{(x,\tau)\in\Omega_T:x\le\overline{x}_s(\tau)\right\}$$

and

$$\left\{(x,\tau)\in\Omega_T:\overline{v}(x,\tau)=\frac{1}{x+1+\lambda}\right\}=\left\{(x,\tau)\in\Omega_T:x\ge\overline{x}_b(\tau)\right\}.$$

Dai and Yi (2009) completely characterized the behaviors of $\overline{x}_s(\tau)$ and $\overline{x}_b(\tau)$, which are summarized as follows.

PROPOSITION 4.1. *Let $\overline{x}_s(\tau)$ and $\overline{x}_b(\tau)$ be two free boundaries as given above in the no-consumption case. Define*

$$\tau_0=\frac{1}{\alpha-r}\log\frac{1+\lambda}{1-\mu}\ \textit{and}\ \tau_1=\frac{1}{\alpha-r-(1-\gamma)\,\sigma^2}\log\frac{1+\lambda}{1-\mu}. \tag{4.2}$$

Then $\overline{x}_s(\tau)<\overline{x}_b(\tau)$, and

(i) *both $\overline{x}_s(\tau)$ and $\overline{x}_b(\tau)$ are monotonically decreasing;*

(ii) *for any $\tau>0$,*

$$-(1-\mu)<\lim_{\tau\to+\infty}\overline{x}_s(\tau)\le\overline{x}_s(\tau)\le\overline{x}_s(0^+)=(1-\mu)\,x_M; \tag{4.3}$$

moreover,

$$\overline{x}_s(\tau)\equiv0\qquad\textit{if }\alpha-r-(1-\gamma)\,\sigma^2=0, \tag{4.4}$$

$$\overline{x}_s(\tau)<0\qquad\textit{if }\alpha-r-(1-\gamma)\,\sigma^2>0, \tag{4.5}$$

$$\overline{x}_s(\tau)>0\qquad\textit{if }\alpha-r-(1-\gamma)\,\sigma^2<0; \tag{4.6}$$

(iii) *for any $\tau>0$,*

$$\overline{x}_b(\tau)\ge(1+\lambda)\,x_M, \tag{4.7}$$

and

$$\overline{x}_b(\tau)=+\infty\ \textit{if and only if}\ \tau\in(0,\tau_0]; \tag{4.8}$$

moreover, when $\alpha-r-(1-\gamma)\,\sigma^2>0$,

$$\overline{x}_b(\tau)>0\ \textit{for}\ \tau<\tau_1,\quad\overline{x}_b(\tau_1)=0,\quad\overline{x}_b(\tau)<0\ \textit{for}\ \tau>\tau_1. \tag{4.9}$$

Remark 4.2. Liu and Loewenstein (2002) obtained partial results of the above proposition, including (4.3), (4.4), (4.7), and (4.8).

4.2. With consumption. It is worthwhile pointing out that $\partial_\tau\overline{v}\le0$ in the no-consumption case, which plays an important role in the analysis of the no-consumption case. But it is not true in the consumption case. Fortunately, we have the following theorem which enables us to extend most results in Proposition 4.1 to the consumption case.

THEOREM 4.3. *Let $x_s(\tau)$ and $x_b(\tau)$ be two free boundaries from problem* (1.10) *(i.e., the consumption case), and let $\overline{x}_s(\tau)$ and $\overline{x}_b(\tau)$ be two free boundaries from problem* (4.1) *(i.e., the no-consumption case). Then*

$$x_s(\tau) \geq \overline{x}_s(\tau), \tag{4.10}$$

$$x_b(\tau) \geq \overline{x}_b(\tau). \tag{4.11}$$

Proof. Let $v(x,\tau)$ and $\overline{v}(x,\tau)$ be the solution to problems (1.10) and (4.1), respectively. By (5.3), we have $\mathcal{L}_w v \leq 0$. It follows from the maximum principle (cf. Friedman (1982), p. 74) that

$$v(x,\tau) \geq \overline{v}(x,\tau).$$

It follows that

$$\begin{aligned} v(x,\tau) &> \frac{1}{x+1+\lambda} \quad \text{if } \overline{v}(x,\tau) > \frac{1}{x+1+\lambda}, \\ \overline{v}(x,\tau) &< \frac{1}{x+1-\mu} \quad \text{if } v(x,\tau) < \frac{1}{x+1-\mu}, \end{aligned}$$

which yields the desired result. □

Remark 4.4. The intuition behind the above theorem is that to maintain consumption, the investor prefers to keep a larger fraction of wealth in the bank account.

Remark 4.5. By (4.3) and (4.10), we infer $x_s(\tau) \geq \overline{x}_s(\tau) > \lim_{\tau\to+\infty} \overline{x}_s(\tau) > -(1-\mu)$. Then we can choose $x^* = \lim_{\tau\to+\infty} \overline{x}_s(\tau)$ such that $x_s(\tau)$ never hits the line $x = x^*$.

THEOREM 4.6. *Let $x_s(\tau)$ and $x_b(\tau)$ be the two free boundaries from problem* (1.10) *(i.e., the consumption case), and let τ_0 be as defined in* (4.2). *Then $x_s(\tau) < x_b(\tau)$, and*

(i) *for any $\tau > 0$,*

$$x_s(\tau) \leq x_s(0^+) \equiv \lim_{\tau\to 0^+} x_s(\tau) = (1-\mu)x_M; \tag{4.12}$$

moreover,

$$x_s(\tau) \equiv 0 \quad \textit{if } \alpha - r - (1-\gamma)\sigma^2 = 0, \tag{4.13}$$

$$x_s(\tau) < 0 \quad \textit{if } \alpha - r - (1-\gamma)\sigma^2 > 0, \tag{4.14}$$

$$x_s(\tau) > 0 \quad \textit{if } \alpha - r - (1-\gamma)\sigma^2 < 0; \tag{4.15}$$

(ii) *for any $\tau > 0$,*

$$x_b(\tau) \geq (1+\lambda)x_M \tag{4.16}$$

and

$$x_b(\tau) = +\infty \ \textit{if and only if } \tau \in (0,\tau_0]. \tag{4.17}$$

Proof. $x_s(\tau) < x_b(\tau)$ is clear. Equation (4.12) has been proved in Lemma 2.4. If $\alpha - r - (1-\gamma)\sigma^2 > 0$, then $x_M < 0$, and (4.14) is a consequence of (4.12). When $\alpha - r - (1-\gamma)\sigma^2 < 0$, (4.15) follows from (4.5) and (4.10). When $\alpha - r - (1-\gamma)\sigma^2 = 0$, we again use (4.5) and (4.10) to get $x_s(\tau) \geq 0$, while we have by (4.12) $x_s(\tau) \leq 0$. This yields (4.13).

The proof of (4.16) is similar to that of (4.12). It can also be obtained from (4.5) and (4.11). As for (4.17), from (4.8) and (4.10) we immediately achieve the sufficiency, namely, $x_b(\tau) = +\infty$ if $\tau \in (0, \tau_0]$. To show the necessity, we need only prove $x_b(\tau) < +\infty$ if $\tau > \tau_0$. By the transformation

$$z = \frac{x}{x+1+\lambda},\ \widetilde{v}(z,\tau) = \left(v(x,\tau) - \frac{1}{x+1+\lambda}\right)\frac{(x+1+\lambda)^2}{1+\lambda}, \tag{4.18}$$

problem (1.10) becomes

$$\begin{cases} \max\left\{\min\left\{\partial_\tau \widetilde{v} - \widetilde{\mathcal{L}}\widetilde{v} + \widetilde{\mathcal{L}}_w\widetilde{v}, \widetilde{v}\right\}, \widetilde{v} - \frac{\lambda+\mu}{(1-\mu)+(\lambda+\mu)z}\right\} = 0, \\ \widetilde{v}(z,0) = \frac{\lambda+\mu}{(1-\mu)+(\lambda+\mu)z} \end{cases} \tag{4.19}$$

in $\frac{1-\mu}{\lambda+\mu} < z < 1$, $\tau > 0$. Here

$$\begin{aligned} \widetilde{\mathcal{L}}\widetilde{v} = {} & \frac{1}{2}\sigma^2 z^2(1-z)^2\partial_{zz}\widetilde{v} - \left((\alpha - r - (2-\gamma)\sigma^2) + 3\sigma^2 z\right) z(1-z)\partial_z\widetilde{v} \\ & - \left(\alpha - r - (1-\gamma)\sigma^2 - 2(\alpha - r - (2-\gamma)\sigma^2)z - 3\sigma^2 z^2\right)\widetilde{v} \\ & - \left(\sigma^2 z + \alpha - r - (1-\gamma)\sigma^2\right) + \gamma\sigma^2 z[1 + (1-z)\widetilde{v}]\left[z(1-z)\partial_z\widetilde{v} + (1-2z)\widetilde{v} + 1\right] \end{aligned}$$

and

$$\widetilde{\mathcal{L}}_w\widetilde{v} = \left(e^{\gamma w}\left[1 + (1-z)\widetilde{v}\right]\frac{1-z}{1+\lambda}\right)^{-\frac{1}{1-\gamma}}\frac{(1-z)^2}{1+\lambda}\left(\widetilde{v}^2 + \partial_z\widetilde{v}\right).$$

Define

$$z_b(\tau) = \sup_z\left\{z \in \left(\frac{1-\mu}{\lambda+\mu}, 1\right) : \widetilde{v}(z,\tau) > 0\right\}.$$

Clearly $z_b(\tau) = \frac{x_b(\tau)}{x_b(\tau)+1+\lambda}$. It suffices to show $z_b(\tau) < 1$ if $\tau > \tau_0$. Noticing $e^{\gamma w} = \gamma V$, it is not hard to verify that $\widetilde{\mathcal{L}}_w\widetilde{v}|_{z=1} = 0$. Therefore, at $z = 1$, problem (4.19) is reduced to

$$\begin{cases} \max\left\{\min\left\{\partial_\tau\widetilde{v}(1,\tau) - (\alpha - r)\,\widetilde{v}(1,\tau) + \alpha - r, \widetilde{v}\right\}, \widetilde{v} - \frac{\lambda+\mu}{1+\lambda}\right\} = 0,\ \tau > 0, \\ \widetilde{v}(1,0) = \frac{\lambda+\mu}{1+\lambda}, \end{cases}$$

whose solution is

$$\widetilde{v}(1,\tau) = \max\left(1 - e^{(\alpha-r)\tau}\frac{1-\mu}{1+\lambda}, 0\right) = \begin{cases} 1 - e^{(\alpha-r)\tau}\frac{1-\mu}{1+\lambda} \text{ when } \tau \in (0,\tau_0], \\ 0 \text{ when } \tau > \tau_0, \end{cases}$$

which yields the desired result. The proof is complete. □

Remark 4.7. Compared with the no-consumption case, the monotonicity of free boundaries is not available for we no longer have $\partial_\tau v \le 0$. A numerical example about the nonmonotonicity is presented in Dai and Zhong (2008). In addition, (4.9) means that in the no-consumption case, $x_b(\tau)$ intersects with $x = 0$ at τ_1, which does not hold in the consumption case because of the additional term $\mathcal{L}_w v$ caused by the consumption. All theoretical results in this section are numerically demonstrated by Dai and Zhong (2008).

Remark 4.8. In the absence of transaction costs, Merton (1971) shows that an investor should never leverage if risk premium $\alpha - r - (1-\gamma)\sigma^2$ is nonpositive. This result remains true in the presence of transaction costs. Indeed, from (4.13)–(4.15), we infer that $x_b \geq x_s \geq 0$ if and only if $\alpha - r - (1-\gamma)\sigma^2 \leq 0$, which implies the conclusion.

Remark 4.9. Equation (4.17) indicates that there is a critical time after which it is never optimal to purchase stocks. This is one interesting and important feature of the finite horizon problem. Its counterpart (4.8) in the no-consumption case was first found by Liu and Loewenstein (2002). The intuition behind this is the following. If the investor does not have a long enough expected horizon to recover at least the transaction costs, then s/he should not purchase any additional stock.

5. Equivalence of Problems A and B. This section is devoted to the equivalence between Problems A and B. Let $(v, w, x_s(\tau))$ be the solution to Problem B. Since Problem A has a unique *viscosity* solution, we need only show that w in Proposition 3.1 is the solution to Problem A.

Define

$$\mathbf{SR} = \left\{ (x,\tau) \in \Omega_T : v(x,\tau) = \frac{1}{x+1-\mu} \right\},$$

$$\mathbf{BR} = \left\{ (x,\tau) \in \Omega_T : v(x,\tau) = \frac{1}{x+1+\lambda} \right\},$$

$$\mathbf{NT} = \left\{ (x,\tau) \in \Omega_T : \frac{1}{x+1+\lambda} < v(x,\tau) < \frac{1}{x+1-\mu} \right\}.$$

In finance, the three regions defined above stand for the selling region, buying region, and no-transaction region, respectively. Due to Proposition 3.1, we already have

$$\mathbf{SR} = \left\{ (x,\tau) \in \Omega_T : x \leq x_s(\tau) \right\}. \tag{5.1}$$

Similar to (2.22)–(2.23), we infer that there is a function $x_b(\tau) : (0,T) \to [x^*, +\infty) \cup +\infty$ such that

$$\mathbf{BR} = \{ (x,t) \in \Omega_T : x \geq x_b(\tau) \}. \tag{5.2}$$

Note that $v = w_x$ satisfies (1.10). Owing to (1.8), we have

$$\frac{\partial}{\partial x}\left(\partial_\tau w - \mathcal{L}_2 w - \frac{1-\gamma}{\gamma} e^{-\frac{\gamma}{1-\gamma}w} \left(\partial_x w\right)^{-\frac{\gamma}{1-\gamma}} \right) \leq 0, \ w_x = \frac{1}{x+1-\mu} \quad \text{in } x \leq x_s(\tau),$$

$$\frac{\partial}{\partial x}\left(\partial_\tau w - \mathcal{L}_2 w - \frac{1-\gamma}{\gamma} e^{-\frac{\gamma}{1-\gamma}w} \left(\partial_x w\right)^{-\frac{\gamma}{1-\gamma}} \right) = 0 \quad \text{in } x_s(\tau) < x < x_b(\tau),$$

$$\frac{\partial}{\partial x}\left(\partial_\tau w - \mathcal{L}_2 w - \frac{1-\gamma}{\gamma} e^{-\frac{\gamma}{1-\gamma}w} \left(\partial_x w\right)^{-\frac{\gamma}{1-\gamma}} \right) \geq 0, \ w_x = \frac{1}{x+1+\lambda} \quad \text{in } x \geq x_b(\tau).$$

Noticing $\partial_\tau w - \mathcal{L}_2 w - \frac{1-\gamma}{\gamma} e^{-\frac{\gamma}{1-\gamma}w} \left(\partial_x w\right)^{-\frac{\gamma}{1-\gamma}} \big|_{x=x_s(\tau)} = 0$, we deduce that w is the solution to Problem A. In terms of Lemma 2.3 and Proposition 3.1, we achieve the following theorem.

THEOREM 5.1. *Problem* A *has a solution* $w(x,\tau) \in W^{2,1}_{\infty}(\Omega_T^R)$ *for any* $R > 0$ *with* $\partial_x w \in C(\overline{\Omega}_T)$ *and* $\partial_{xx} w, \partial_\tau w \in L^\infty(\Omega_T) \cap C\left(\overline{\Omega}_T \backslash \{x = 0\}\right)$. *Moreover,* $v = \partial_x w$ *satisfies problem* (1.10),

$$-\frac{K}{(x+1-\mu)^2} \leq \partial_x v \leq -v^2, \tag{5.3}$$

$$|\partial_\tau w| \leq M,$$

where K *and* M *are positive constants.*

6. Conclusion.

In this paper, we study the optimal investment and consumption decision of a CRRA investor who faces proportional transaction costs and a finite time horizon. Most of the previous work takes only either an infinite time horizon or pure investment without consumption into consideration.

Mathematically the problem can be formulated as a singular stochastic control problem. It turns out that the value function is governed by a degenerate parabolic variational inequality with gradient constraints, which gives rise to two free boundaries. We aim to investigate the behaviors of the free boundaries which, respectively, stand for the optimal buying and selling strategies.

Since it is intractable to study the free boundaries directly from the original variational inequality with gradient constraints, following Dai and Yi (2009) which dealt with the no-consumption case, we manage to derive a standard variational inequality (i.e., an obstacle problem) that some partial derivative of the value function satisfies. In terms of the later variational inequality, it is rather straightforward to characterize the behaviors of the free boundaries and to study the regularity of the value function. In essence, our approach relies on the connection between singular control and optimal stopping, which is, though well known in the field of singular stochastic control, never revealed for the present problem. Our approach can also be utilized to handle the infinite horizon problems.

Compared with the no-consumption case, the free boundaries are no longer monotone, but most of the other results remain valid. For instance, there is a critical time after which it is never optimal to purchase stocks. The no-trading region is always in the first quadrant if and only if $\alpha - r - (1-\gamma)\sigma^2 \le 0$, which means that leverage is always suboptimal if risk premium is nonpositive.

It is worthwhile pointing out that a technical condition $\gamma > 0$ is required (see Remark 2.2). We believe it could be removed and would like to leave this for future research.

Appendix A. The case of logarithm utility function ($\gamma = 0$).

In this case $U(W) = \ln(W)$. Then the differential operator $\mathscr{L}$ is replaced by

$$\mathscr{L}\varphi = \frac{1}{2}\sigma^2 y^2 \partial_{yy}\varphi + \alpha y \partial_y \varphi + r x \partial_x \varphi - \beta\varphi - (1 + \ln(\partial_x \varphi)).$$

The homotheticity property of the utility function leads to

$$\varphi(\rho x, \rho y, t) = g(t)\ln\rho + \varphi(x,y,t),$$

where $g(t) = \frac{1-e^{-\beta(T-t)}}{\beta} + e^{-\beta(T-t)}$. This motivates us to make the transformation

$$V\left(\frac{x}{y}, \tau\right) = g(t)\ln\frac{1}{y} + \varphi(x,y,t).$$

It follows that

$$\begin{cases} \min\{\partial_\tau V - \mathcal{L}_0 V, -(x+1-\mu)\partial_x V + \bar{g}(\tau), (x+1+\lambda)\partial_x V - \bar{g}(\tau)\} = 0, \\ V(x,0) = \ln(x+1-\mu), \qquad -(1-\mu) < x < +\infty,\ 0 < \tau \le T, \end{cases}$$

where $\bar{g}(\tau) = \frac{1-e^{-\beta\tau}}{\beta} + e^{-\beta\tau}$.

$$\mathcal{L}_0 V = \frac{1}{2}\sigma^2 x^2 \partial_{xx} V - \left(\alpha - r - \sigma^2\right) x \partial_x V - \beta V + \left(\alpha - \frac{1}{2}\sigma^2\right)\bar{g}(\tau) - (1 + \ln(\partial_x V)).$$

Let $w(x,\tau) = \frac{V(x,\tau)}{\bar{g}(\tau)}$. It then follows that

$$\begin{cases} \min\left\{\partial_\tau w - \mathcal{L}_3 w, -(x+1-\mu)\partial_x w + 1, (x+1+\lambda)\partial_x w - 1\right\} = 0, \\ w(x,0) = \ln(x+1-\mu), \qquad\qquad -(1-\mu) < x < +\infty,\ 0 < \tau \le T, \end{cases}$$

where

$$\mathcal{L}_3 w = \frac{1}{2}\sigma^2 x^2 \partial_{xx} w - \left(\alpha - r - \sigma^2\right) x \partial_x w + \left(\alpha - \frac{1}{2}\sigma^2\right) - \frac{1}{\bar{g}(\tau)}\left(1 + \ln \bar{g}(\tau) + w + \ln \partial_x w\right).$$

Set

$$v = \partial_x w.$$

We postulate that v is the solution to the following double obstacle problem:

$$\text{(A.1)} \qquad \begin{cases} \partial_\tau v - \mathcal{L}_4 v = 0 & \text{if } \frac{1}{x+1+\lambda} < v < \frac{1}{x+1-\mu}, \\ \partial_\tau v - \mathcal{L}_4 v \le 0 & \text{if } v = \frac{1}{x+1-\mu}, \\ \partial_\tau v - \mathcal{L}_4 v \ge 0 & \text{if } v = \frac{1}{x+1+\lambda}, \\ v(x,0) = \frac{1}{x+1-\mu}, & -(1-\mu) < x < +\infty,\ 0 < \tau \le T, \end{cases}$$

where

$$\mathcal{L}_4 v = \frac{1}{2}\sigma^2 x^2 \partial_{xx} v - \left(\alpha - r - 2\sigma^2\right) x \partial_x v - \left(\alpha - r - \sigma^2\right) v - \frac{1}{\bar{g}(\tau)}\left(v + \frac{\partial_x v}{v}\right).$$

Note that $\mathcal{L}_4 v$ is independent of w. To study the problem, we can adopt a similar argument as in Dai and Yi (2009) and the same treatment on degeneracy used in the present paper. The details are omitted.

Appendix B. On the case of $y < 0$. The reduction of dimension in (1.6) is confined to the case of $y > 0$. To extend to the case of $y < 0$, we can make another transformation. Indeed, due to $x + (1+\lambda)\, y > 0$, we get by the homotheticity

$$\begin{aligned} \frac{1}{\left(x + (1+\lambda)\, y\right)^\gamma}\varphi(x,y,t) &= \varphi\left(\frac{x}{x+(1+\lambda)\, y}, \frac{y}{x+(1+\lambda)\, y}, t\right) \\ &= \varphi\left(z, \frac{1-z}{1+\lambda}, t\right) \equiv \widetilde{V}(z,\tau), \end{aligned}$$

where $z = \frac{x}{x+(1+\lambda)y}$. This reduction applies to the case of $y < 0$. Similar to the previous arguments, we define $\widetilde{v}(z,\tau) = \frac{\partial_z \widetilde{V}}{\gamma \widetilde{V}}$. In contrast to $V(x,\tau)$ and $v(x,\tau)$, it is not hard to verify that $z = \frac{x}{x+1+\lambda}$,

$$\frac{1}{(x+1+\lambda)^\gamma} V(x,\tau) = \widetilde{V}\left(z,\tau\right), \text{ and } \widetilde{v}(z,\tau) = \left(v(x,\tau) - \frac{1}{x+1+\lambda}\right)\frac{(x+1+\lambda)^2}{1+\lambda}$$

for $z < 1$. This is nothing but the transformation (4.18) which yields problem (4.19). We emphasize that problem (4.19) in $z > 1$ corresponds to the case of $y < 0$ subject to the initial condition $\widetilde{v}(z,0) = 0$ (keep in mind that $\varphi(x,y,T) = \frac{1}{\gamma}\left(x + (1+\lambda)\, y\right)^\gamma$ for $y < 0$). Then, it is easy to see that $\widetilde{v}(z,\tau) \equiv 0$ is the *unique* solution to problem (4.19) in $z > 1$, $\tau \in (0,T]$ because

$$\partial_\tau 0 - \widetilde{\mathcal{L}} 0 + \widetilde{\mathcal{L}}_w 0 = \alpha - r > 0.$$

This implies that the whole region $\{z > 1\}$ (i.e., $\{y < 0\}$) is in the buy region, that is, leverage is never optimal for $\alpha > r$. It is worthwhile pointing out that $\widetilde{v}(z, \tau)$ is likely to be discontinuous across $z = 1$, but the value function $\varphi(x, y, t)$ must be continuous across $y = 0$.

REFERENCES

M. Akian, J. L. Menaldi, and A. Sulem (1996), *On an investment-consumption model with transaction costs*, SIAM J. Control Optim., 34, pp. 329–364.

J. Cvitanić and I. Karatzas (1996), *Hedging and portfolio optimization under transaction costs: A martingale approach*, Math. Finance, 6, pp. 133–165.

M. Dai, Z. Q. Xu, and X. Y. Zhou (2007), *Continuous-Time Mean-Variance Portfolio Selection with Proportional Transaction Costs*, Working paper, National University of Singapore, Singapore.

M. Dai and F. H. Yi (2009), *Finite horizon optimal investment with transaction costs: A parabolic double obstacle problem*, J. Differential Equations, 246, pp. 1445–1469.

M. Dai and Y. F. Zhong (2008), *Penalty Methods for Continuous-Time Portfolio Selection with Proportional Transaction Costs*, Working paper, National University of Singapore, Singapore.

M. H. A. Davis and A. R. Norman (1990), *Portfolio selection with transaction costs*, Math. Oper. Res., 15, pp. 676–713.

M. H. A. Davis, V. G. Panas, and T. Zariphopoulou (1993), *European option pricing with transaction costs*, SIAM J. Control Optim., 31, pp. 470–493.

L. C. Evans (1979), *A second-order elliptic equation with gradient constraint*, Comm. Partial Differential Equations, 4, pp. 555–572.

W. H. Fleming and H. M. Soner (2006), *Controlled Markov Processes and Viscosity Solutions*, 2nd ed., Springer-Verlag, New York.

A. Friedman (1982), *Variational Principles and Free-Boundary Problems*, Wiley, New York.

B. Hu (1986), *Fully nonlinear elliptic equations with gradient constraint*, Beijing Daxue Xuebao, 5, pp. 78–91.

H. Ishii and S. Koike (1983), *Boundary regularity and uniqueness for an elliptic equation with gradient constraint*, Comm. Partial Differential Equations, 8, pp. 317–346.

K. Janeček and S. Shreve (2004), *Asymptotic analysis for optimal investment and consumption with transaction costs*, Finance Stoch., 8, pp. 181–206.

I. Karatzas and S. E. Shreve (1984), *Connections between optimal stopping and singular stochastic control* I: *Monotone follower problems*, SIAM J. Control Optim., 22, pp. 856–877.

I. Karatzas and S. E. Shreve (1998), *Methods of Mathematical Finance*, Springer-Verlag, New York.

T. L. Lai and T. W. Lim (2003), *Singular stochastic control in optimal investment and hedging in the presence of transaction costs*, in Probability, Statistics, and Their Applications, IMS Lecture Notes Monogr. Ser. 41, Inst. Math. Statist., Beachwood, OH, pp. 209–227.

H. Liu and M. Loewenstein (2002), *Optimal portfolio selection with transaction costs and finite horizons*, Review of Financial Studies, 15, pp. 805–835.

M. J. P. Magill and G. M. Constantinides (1976), *Portfolio selection with transaction costs*, J. Econom. Theory, 13, pp. 264–271.

H. M. Markowitz (1952), *Portfolio selection*, J. Finance, 7, pp. 77–91.

R. C. Merton (1969), *Lifetime portfolio selection under uncertainty: The continuous-time case*, Rev. Econom. Statist., 51, pp. 247–257.

R. C. Merton (1971), *Optimum consumption and portfolio rules in a continuous-time model*, J. Econom. Theory, 3, pp. 373–413.

S. E. Shreve and H. M. Soner (1994), *Optimal investment and consumption with transaction costs*, Ann. Appl. Probab., 4, pp. 609–692.

H. M. Soner and S. E. Shreve (1991), *A free boundary problem related to singular stochastic control: The parabolic case*, Comm. Partial Differential Equations, 16, pp. 373–424.

M. Wiegner (1981), *The $C^{1,1}$-character of solutions of second order elliptic equations with gradient constraint*, Comm. Partial Differential Equations, 6, pp. 361–371.

J. M. Yong and X. Y. Zhou (1999), *Stochastic Controls: Hamiltonian Systems and HJB Equations*, Springer-Verlag, New York.

H. Zhu (1992), *Generalized solution in singular stochastic control: The nondegenerate problem*, Appl. Math. Optim., 25, pp. 225–245.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 2, pp. 1155–1178

© 2009 Society for Industrial and Applied Mathematics

UNKNOWN INPUT AND STATE ESTIMATION FOR UNOBSERVABLE SYSTEMS*

FRANCISCO J. BEJARANO[†], LEONID FRIDMAN[†], AND ALEXANDER POZNYAK[‡]

Abstract. The concept of strong detectability and its relation with the concept of invariant zeros is reviewed. For strongly detectable systems (which includes the strongly observable systems), it is proposed a hierarchical design of a robust observer whose trajectories converge to those of the original state vector. Furthermore, it is shown that neither left invertibility is a sufficient condition nor strong detectability is a necessary condition to estimate the unknown inputs. It is shown that the necessary and sufficient condition for estimating the unknown inputs is that the set of the invariant zeros that do not belong to the set of unobservable modes be within the interior of the left half plane of the complex space. This shows that the unknown inputs could be estimated even if it is impossible to estimate the entire state vector of the system. Two numerical examples illustrate the effectiveness of the proposed estimation schemes.

Key words. unknown input estimation, strong detectability, sliding mode observer

AMS subject classifications. 93C41, 93B07, 93B51

DOI. 10.1137/070700322

1. Introduction.

1.1. Antecedents.

The problem of state observation for systems with unknown inputs has been extensively studied in the last two decades. Usually, the design of observers requires the system to have relative degree one with respect to the unknown inputs (see, e.g., [16] and [10]). Within *variable structure theory,* the problems of state observation and unknown input estimation have been actively developed using the *sliding mode* approach (see, for example, the corresponding chapters in the textbooks [11], [27] and the recent tutorials [3], [10], [22]). But generally they were developed for systems which satisfy the necessary and sufficient conditions to estimate the entire state vector without differentiation of the output (i.e., for the systems with relative degree one w.r.t. the unknown inputs) [16]. It turns out that the previously mentioned conditions are not satisfied for the state observation of a mechanical system with sensors measuring only the position of the elements of the system [9].

To overcome the restriction of relative degree one w.r.t. the unknown inputs, an idea was suggested: to transform the system into a triangular form and use a step-by-step sliding mode observer based on the successive reconstruction of each element of the transformed state vector (see, e.g., [15], [27], [1], and [14]). However, the design of those observers is restricted to the fulfilment of a specific relative degree condition ([12]). The essence of the observers that use the triangular form is to recover information from the derivatives of the output of the system which are not affected by the unknown inputs. Such derivatives can be estimated via a second-order sliding

*Received by the editors August 16, 2007; accepted for publication (in revised form) December 3, 2008; published electronically March 18, 2009. Results of this manuscript were presented in the European Control Conference 2007, Kos, Greece.
http://www.siam.org/journals/sicon/48-2/70032.html

[†]Department of Control, Division of Electrical Engineering, National Autonomous University of Mexico (UNAM), C.P. 04510, México, D.F. (javbejarano@yahoo.com.mx, lfridman@servidor.unam.mx).

[‡]Departamento de Control Automático, CINVESTAV-IPN, A.P. 14-740, C.P. 07000 México D.F. (apoznyak@ctrl.cinvestav.mx).

mode technique, specifically by the super-twisting algorithm. In the last two decades some second-order sliding-mode algorithms have been designed (see, e.g., [2], [4], [24], [5], and [21]). The super-twisting technique is a second-order sliding mode that keeps the advantages of classic sliding mode, and further the super-twisting algorithm can be used as a *robust exact differentiator* [18, 19]. It is used here for the state and unknown estimation.

1.2. Motivation. It was shown in [7, 8] that strong observability condition (absence of invariant zeros) is necessary and sufficient for the reconstruction in finite time of the state vector. Regarding the observation problem, in this paper we suggest a scheme of design which relaxes the strong observability condition, even when the convergence of the observation error to the zero point becomes asymptotic.

On the other hand, usually the estimation of the unknown inputs requires first to estimate the entire state vector (see, e.g., [25], [23], [13]); however, estimating the entire state vector, as we shall see below, requires the system to be at least strongly detectable (equivalently, that the set of the invariant zeros belongs to the interior of the left half plane of the complex space). Here, we show that in the general case, for the unknown input estimation, the strong detectability condition can be relaxed.

1.3. Main contributions. Regarding the unknown input reconstruction, the main contributions of this paper are:

- Necessary and sufficient structural conditions for the unknown input estimation have been found. Namely, the estimation of the unknown inputs can be carried out if the set of the invariant zeros of the system (for the known control input equal to zero) that do not belong to the set of unobservable eigenvalues is within the interior of the left half plane of the complex space.
- The structural conditions under which the unknown inputs can be reconstructed exactly in a finite time are given.
- A scheme for the estimation (reconstruction) of the unknown inputs is suggested, which is based on the decomposition of the system into three subsystems. This allows one to estimate the states of the first two subsystem, which is enough for the unknown input estimation (reconstruction).
- Combining the structural conditions obtained in this paper and the conditions given in [16] for state estimation, it is shown that, under more restrictive conditions, the unknown inputs could be estimated without estimating the entire state vector and without using any derivative of the system output.

If the system is not strongly observable, it is impossible to design a standard differential observer providing state estimation. Because of that, we proposed another approach related to the designing of an algebraic-type observer which can successfully work for both strongly observable and nonstrongly observable, but strongly detectable systems. Therefore, concerning the state estimate, the main contributions are:

- Decomposition of the system into two subsystems. The first one is strongly observable for the null known control input and the second one is expected to be detectable.
- Design of an observer for the state vector of the first subsystem by applying the *second-order hierarchical observation* scheme [7].

1.4. Structure of the paper. The manuscript is structured in the following manner. In section 2 we outline the problem statement. Section 3 is devoted to some preliminaries dealing mainly with the concepts of strong observability and strong detectability. In the same section, we present the main idea for estimating the state vector. Necessary and sufficient conditions under which the unknown inputs can be estimated are given in section 4. Section 5 deals with the design of the observer of the state vector. In section 6, an algorithm for the estimation of the unknown inputs is suggested. Some simulations are depicted in section 7, which illustrate the scheme of design proposed in the paper. The proofs of propositions, lemmas, and theorems are given in the appendix.

1.5. Notation. We use the following notation. Let $G \in \mathbb{R}^{n\times m}$ be a matrix. We define G^{+} as the pseudoinverse of G. Thus, if $\operatorname{rank} G = n$, $GG^{+} = I$, and if $\operatorname{rank} G = m$, $G^{+}G = I$. For $J \in \mathbb{R}^{n\times m}$ with $\operatorname{rank} J = r$, we define $J^{\perp} \in \mathbb{R}^{n-r\times n}$ with $\operatorname{rank} J^{\perp} = n - r$ as a matrix achieving $J^{\perp}J = 0$; and $J^{\perp\perp} \in \mathbb{R}^{r\times n}$ with $\operatorname{rank} J^{\perp\perp} = r$ as a matrix such that $J^{\perp}(J^{\perp\perp})^{T} = 0$. Notice that $\det\left[\begin{smallmatrix} J^{\perp} \\ J^{\perp\perp}\end{smallmatrix}\right] \neq 0$, and also that $J^{\perp\perp}J \in \mathbb{R}^{r\times m}$ and $\operatorname{rank}(J^{\perp\perp}J) = r$. $\mathbb{C}^{-} := \{s \in \mathbb{C} : \operatorname{Re} s < 0\}$.

2. Problem formulation. Let us consider the following system affected by unknown inputs:[1]

$$\begin{array}{l} \dot{x}\,(t) = Ax\,(t) + Bu\,(t) + Dw\,(t)\,,\ x\,(0) = x_0 \\ y\,(t) = Cx\,(t) + Fw\,(t)\,,\ t \geq 0 \end{array} \tag{2.1}$$

The vector $x\,(t) \in \mathbb{R}^{n}$ is the state vector, $u\,(t) \in \mathbb{R}^{m}$ is the control, $y\,(t) \in \mathbb{R}^{p}$ is the output of the system, $w\,(t) \in \mathbb{R}^{q}$ represents the unknown input vector, which is bounded, i.e., $\|w\,(t)\| \leq w^{+} < \infty$. The matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, $D \in \mathbb{R}^{n\times q}$, and $F \in \mathbb{R}^{p\times q}$ are known constants. The pair $\{u\,(t)\,, y\,(t)\}$ is assumed to be measurable (available) at any time $t \geq 0$. The current states $x\,(t)$ as well as the initial state x_0 are not available. Without the loss of generality we assume that

$$\operatorname{rank}\begin{bmatrix} D \\ F \end{bmatrix} = q.$$

Problems statement:

In this paper we would like to discuss the following problems for the system (2.1):

(a) the estimation of $x\,(t)$ based on the available information $\{u\,(\tau)\,, y\,(\tau)\}_{\tau\in[0,t]}$,

(b) the estimation of $w\,(t)$ based on the available information $\{u\,(\tau)\,, y\,(\tau)\}_{\tau\in[0,t]}$.

3. Preliminaries. Defining $\dot{x}_{\mathrm{c}} = Ax_{\mathrm{c}}\,(t) + Bu$ we have that the dynamic equation for $x_{\mathrm{e}} := x - x_{\mathrm{c}}$ is given by $\dot{x}_{\mathrm{e}}\,(t) = Ax_{\mathrm{e}}\,(t) + Dw\,(t)$ with the output $y_{\mathrm{e}} := y - Cx_{\mathrm{c}} = Cx_{\mathrm{e}}\,(t) + Fw\,(t)$. Thus, the estimation of x is equivalent to the estimation of x_{e} since $x = x_{\mathrm{e}} + x_{\mathrm{c}}$. It means that for the observation problem the control u does not play any role. Therefore, without the loss of generality it will be assumed throughout this section and the next one that $u \equiv 0$.

Let $\Sigma := (A, C, D, F)$ be the fourfold of matrices associated to the dynamic system that is governed by the equations

$$\begin{array}{l} \dot{x}\,(t) = Ax\,(t) + Dw\,(t)\,,\ x\,(0) = x_0 \\ y\,(t) = Cx\,(t) + Fw\,(t)\,,\ t \geq 0 \end{array} \tag{3.1}$$

3.1. Strong observability. We recall some definitions corresponding to properties of Σ and its associated dynamic system (3.1) (see, e.g., [16], [26]).

[1] It can be done an extension to the case of nonlinear systems considered in [14].

Definition 3.1. *The system* (3.1) *is called* ***strongly observable*** *if, for all initial condition* x_0 *and for all unknown input* $w(t)$, *the identity* $y(t) = 0$ *for all* $t \geq 0$ *implies that* $x(t) = 0$ *for all* $t \geq 0$.

Definition 3.2. $\mathcal{V}_\Sigma$ *is a* ***null-output*** (A, D) ***invariant subspace*** *if for every* $x_0 \in \mathcal{V}$ *there exists a* w *such that* $(Ax_0 + Dw) \in \mathcal{V}_\Sigma$ *and* $(Cx_0 + Fw) = 0$. $\mathcal{V}_\Sigma^*$ *is the maximal null-output* (A, D) *invariant subspace; i.e., for every* $\mathcal{V}_\Sigma$ *we have* $\mathcal{V}_\Sigma \subset \mathcal{V}_\Sigma^*$. *The subspace* $\mathcal{V}_\Sigma^*$ *is called the* ***weakly unobservable subspace*** *of* Σ.

Definition 3.3. $s_0 \in \mathbb{C}$ *is an* ***invariant zero*** *of* Σ *if*

$$\operatorname{rank} \bar{P}(s_0) < n + \operatorname{rank}\begin{bmatrix} D \\ F \end{bmatrix};\ \bar{P}(s_0) := \begin{bmatrix} s_0 I - A & -D \\ C & F \end{bmatrix}. \tag{3.2}$$

Fact 1. *The following statements are equivalent (see, e.g.,* [16], [26]*):*
(i) *The dynamic system* (3.1), *associated to* Σ, *is strongly observable;*
(ii) $\mathcal{V}_\Sigma^* = 0$*;*
(iii) Σ *has no invariant zeros.*

3.2. Decomposition into the strongly and nonstrongly observable subsystems. Now, we will decompose the system into the strongly observable part and the nonstrongly observable part. With this aim, we will need a basis of $\mathcal{V}_\Sigma^*$. Next, we give a form to construct a basis for the subspace $\mathcal{V}_\Sigma^*$. Let the matrices $M_{k,\Sigma}$ be defined recursively by

$$\begin{gathered} M_{k+1,\Sigma} = \bar{M}^{\perp\perp}_{k+1,\Sigma}\bar{M}_{k+1,\Sigma}\ ,\ M_{1,\Sigma} = \left(F^\perp C\right)^{\perp\perp} F^\perp C \\ \bar{M}_{k+1,\Sigma} = T_{k,\Sigma}\begin{pmatrix} M_{k,\Sigma}A \\ C \end{pmatrix},\ T_{k,\Sigma} = \begin{pmatrix} M_{k,\Sigma}D \\ F \end{pmatrix}^{\perp}. \end{gathered} \tag{3.3}$$

Thus, $M_{k+1,\Sigma}$ has full row rank.[2] In [20] it was proven that

$$\mathcal{V}_\Sigma^* = \ker M_{n,\Sigma}. \tag{3.4}$$

Defining[3] $n_1 := \operatorname{rank} M_{n,\Sigma}$, we have that $M_{n,\Sigma} \in \mathbb{R}^{n_1 \times n}$. Now, with $V \in \mathbb{R}^{n \times n - n_1}$ being a matrix whose columns form a basis of $\mathcal{V}_\Sigma^*$, define the following nonsingular matrix

$$P := \begin{bmatrix} M_{n,\Sigma} \\ V^+ \end{bmatrix}, \tag{3.5}$$

where $V^+ \in \mathbb{R}^{n-n_1 \times n}$. Hence,[4] $P^{-1} = \begin{bmatrix} M^+_{n,\Sigma} & V \end{bmatrix}$, $M^+_{n,\Sigma} \in \mathbb{R}^{n \times n_1}$. On the other hand, Definition 3.2 is equivalent to the fulfilment of the following pair of algebraic equations

$$AV + DK^* = VQ,\ CV + FK^* = 0 \tag{3.6}$$

[2] According to the notation given in 1.5, the matrix $\bar{M}^{\perp\perp}_{k+1,\Sigma}$ has full row rank and $\operatorname{rank}(M_{k+1,\Sigma}) = \operatorname{rank}(\bar{M}^{\perp\perp}_{k+1,\Sigma}) = \operatorname{rank}(\bar{M}_{k+1,\Sigma})$. At difference with the definition of M_{k+1} given in [7], here $M_{k+1,\Sigma}$ always has full row rank.

[3] It is easy to verify that $\operatorname{rank} M_{j+1} = \operatorname{rank} M_j$ implies $\operatorname{rank} M_{j+2} = \operatorname{rank} M_j$. Therefore, to reduce the number of computations of the matrices M_k, if $\operatorname{rank} M_{j+1} = \operatorname{rank} M_j$, we can define $M_n = M_{n-1} = \cdots = M_j$.

[4] Notice that $M^+_{n,\Sigma} = M^T_{n,\Sigma}(M_{n,\Sigma}M^T_{n,\Sigma})^{-1}$ and $V^+ = (V^TV)^{-1}V^T$. Therefore, $M_{n,\Sigma}M^+_{n,\Sigma} = I$, $V^+V = I$, $M_{n,\Sigma}V = 0$, $V^+M^+_{n,\Sigma} = 0$.

for some matrices $\{K^*, Q\}$. It is clear that there exists a matrix $\bar{K}^* \in \mathbb{R}^{m\times n}$ such that

$$\left\{\begin{matrix} AV + DK^* = VQ \\ CV + FK^* = 0 \end{matrix}\right\} \overset{\text{equivalent}}{\Longleftrightarrow} \left\{\begin{matrix} \left(A + D\bar{K}^*\right) V = VQ \\ \left(C + F\bar{K}^*\right) V = 0 \end{matrix}\right\}. \tag{3.7}$$

Taking into account that $V^+V = I$, it is easy to see that $\bar{K}^* = K^*V^+$ satisfies (3.7). It also should be noticed that in general K^* is not unique. Let $\bar{x}$ be defined by $\bar{x} = Px$ with the partition $\bar{x}^T = \begin{bmatrix}\bar{x}_1^T & \bar{x}_2^T\end{bmatrix}$, where $\bar{x}_1 \in \mathbb{R}^{n_1}$ and $\bar{x}_2 \in \mathbb{R}^{n-n_1}$. Thus, because of the manner in which P was defined, and from (3.7) and (3.4), the dynamics of $\bar{x}$ is governed by the equations

$$\begin{aligned} \begin{bmatrix}\dot{\bar{x}}_1(t) \\ \dot{\bar{x}}_2(t)\end{bmatrix} &= \begin{bmatrix} A_1 & 0 \\ A_2 & A_4 \end{bmatrix}\begin{bmatrix}\bar{x}_1(t) \\ \bar{x}_2(t)\end{bmatrix} + \begin{bmatrix} D_1 \\ D_2\end{bmatrix}\bar{w}(t) \\ y(t) &= C_1\bar{x}_1(t) + F\bar{w}(t) \\ \bar{w}(t) &= w(t) - \bar{K}^*P^{-1}\bar{x} = w(t) - K^*\bar{x}_2(t), \end{aligned} \tag{3.8}$$

where

$$\begin{bmatrix} A_1 & 0 \\ A_2 & A_4 \end{bmatrix} := P\left(A + D\bar{K}^*\right)P^{-1}, \ \begin{bmatrix} D_1 \\ D_2 \end{bmatrix} := PD, \ C_1 := \left(C + F\bar{K}^*\right)M_{n,\Sigma}^+. \tag{3.9}$$

Now, define

$$\Sigma_{\bar{K}^*,P} := \left(P\left(A + D\bar{K}^*\right)P^{-1}, \left(C + F\bar{K}^*\right)P^{-1}, PD, F\right).$$

From (3.4) and (3.7), it follows that

$$\ker M_{n,\Sigma_{\bar{K}^*,P}} = \mathcal{V}^*_{\Sigma_{\bar{K}^*,P}} = P\mathcal{V}^*_\Sigma = P\ker M_{n,\Sigma}. \tag{3.10}$$

Lemma 3.4. *Defining $\bar{\Sigma} := (A_1, C_1, D_1, F)$, we have that the dynamic system associated to $\bar{\Sigma}$ is strongly observable; i.e., $\ker M_{n_1,\bar{\Sigma}} = 0$ and $\bar{\Sigma}$ has no invariant zeros.*

3.3. Strong detectability.

The next definition can be found in [16] and [26].

Definition 3.5. *The system* (3.1) *is called **strongly detectable** if, for all initial condition x_0 and for all unknown inputs $w(t)$ providing the existence of solution in* (3.1), *the identity $y(t) = 0$ for all $t \geq 0$ implies $x(t) \to 0$ as $t \to \infty$.*

Remark 1. *It is clear that the strong detectability property is a necessary requirement for the asymptotic estimation of the state vector. As we will see later, strong detectability is also a sufficient condition for such a purpose.*

Remark 2. *Evidently, the strong detectability condition is less restrictive than the strong observability condition. The following system is an example of a system that is not strongly observable, but it is strongly detectable.*

$$\begin{aligned} \dot{x}_1(t) &= x_1(t) + w(t) \\ \dot{x}_2(t) &= x_1(t) - x_2(t) \\ y(t) &= x_1(t) \end{aligned}$$

The following theorem relates the strong detectability with the invariant zeros.

Theorem 3.6 (see [16]). *The system* (3.1) *is strongly detectable if, and only if, the set of the invariant zeros of Σ belongs to $\mathbb{C}^-$.*

Now, using the notation (3.9), we are ready to give a characterization of the invariant zeros of Σ.

LEMMA 3.7. *The invariant zeros of* $\Sigma := (A, C, D, F)$ *are characterized by the following properties:*

(a) *If* $\operatorname{rank}\left[\begin{smallmatrix} D_1 \\ F \end{smallmatrix}\right] = q$*, the set of the invariant zeros of* Σ *and the set of eigenvalues of the matrix* A_4 *are identical.*

(b) *If* $\operatorname{rank}\left[\begin{smallmatrix} D_1 \\ F \end{smallmatrix}\right] < q$*, every* $s \in \mathbb{C}$ *is an invariant zero, where* q *is the number of unknown inputs, that is,* $w(t) \in \mathbb{R}^q$*.*

The following proposition can be found as an exercise on p. 170 of [26].

PROPOSITION 3.8. *The system* $\Sigma = (A, C, D, F)$ *is strongly detectable if, and only if, the pair* $(A + DK, C + FK)$ *is detectable for any* $K \in \mathbb{R}^{q \times n}$*.*

3.4. Basic idea for the state estimation.

In this part of the paper, we will give the basic procedure for the reconstruction of the state in the new coordinates.

3.4.1. Recursive method for the reconstruction of $\bar{x}_1$.

The next is a recursive method for expressing $\bar{x}_1$ as a function of y and its derivatives. It consists in the successive construction of the vectors $M_{k,\bar{\Sigma}}\bar{x}_1(t)$, which leads to the construction of the vector $M_{n_1,\bar{\Sigma}}\bar{x}_1(t)$.

Construction of the vector $M_{n_1,\bar{\Sigma}}\bar{x}_1(t)$ $\left(\bar{\Sigma} := A_1, C_1, D_1, F\right)$:

1. Defining $\xi_1(y) := (F^\perp C_1)^{\perp\perp} F^\perp y$, the following equality is obtained

$$\xi_1(y) = \left(F^\perp C_1\right)^{\perp\perp} F^\perp C_1 \bar{x}_1 = M_{1,\bar{\Sigma}}\bar{x}_1;$$

2. defining $\xi_2(y, \dot{y}) := \bar{M}_{2,\bar{\Sigma}}^{\perp\perp} T_{1,\bar{\Sigma}} \left[\begin{smallmatrix} \frac{d}{dt} F^\perp C_1 \bar{x}_1 \\ y \end{smallmatrix}\right]$, it is obtained

$$\xi_2(y, \dot{y}) = \bar{M}_{2,\bar{\Sigma}}^{\perp\perp} T_{1,\bar{\Sigma}} \begin{bmatrix} F^\perp C_1 A_1 \\ C_1 \end{bmatrix} \bar{x}_1 = M_{2,\bar{\Sigma}}\bar{x}_1;$$

$k+1$. defining $\xi_{k+1}(y, \dot{y}, \ldots, y^{(k)}) := \bar{M}_{k+1,\bar{\Sigma}}^{\perp\perp} T_{k,\bar{\Sigma}} \left[\begin{smallmatrix} \frac{d}{dt} M_{k,\bar{\Sigma}} \bar{x}_1 \\ y \end{smallmatrix}\right]$, it is obtained

$$\xi_{k+1}\left(y, \dot{y}, \ldots, y^{(k)}\right) = \bar{M}_{k+1,\bar{\Sigma}}^{\perp\perp} T_{k,\bar{\Sigma}} \begin{bmatrix} M_{k,\bar{\Sigma}} A_1 \\ C_1 \end{bmatrix} \bar{x}_1 = M_{k+1,\bar{\Sigma}}\bar{x}_1;$$

n_1. finally, defining $\xi_{n_1}(y, \dot{y}, \ldots, y^{(n_1-1)}) := \bar{M}_{n_1,\bar{\Sigma}}^{\perp\perp} T_{n_1-1,\bar{\Sigma}} \left[\begin{smallmatrix} \frac{d}{dt} M_{n_1-1,\bar{\Sigma}} \bar{x}_1 \\ y \end{smallmatrix}\right]$, one gets

$$\xi_{n_1}\left(y, \dot{y}, \ldots, y^{(n_1-1)}\right) = \bar{M}_{n_1,\bar{\Sigma}}^{\perp\perp} T_{n_1-1,\bar{\Sigma}} \begin{bmatrix} M_{n_1-1,\bar{\Sigma}} A_1 \\ C_1 \end{bmatrix} \bar{x}_1 = M_{n_1,\bar{\Sigma}}\bar{x}_1.$$

From Lemma 3.4, the equivalence between (ii) and (iii) in Fact 1, and (3.4), we have that $\det M_{n_1,\bar{\Sigma}} \neq 0$. Thus, after premultiplying by $M_{n_1,\bar{\Sigma}}^{-1}$ in the n_1th stage, the vector $\bar{x}_1(t)$ can be expressed by means of the following formula:

$$\bar{x}_1(t) = M_{n_1,\bar{\Sigma}}^{-1} \xi_{n_1}\left(y, \dot{y}, \ldots, y^{(n_1-1)}\right) \tag{3.11}$$

Equation (3.11) means that $\bar{x}_1$ always can be reconstructed by means of linear combinations of the terms of the vector y and their derivatives.

3.4.2. Procedure for the estimation of $\bar{x}_2$. As we have seen below, the strong detectability property is a necessary condition for the asymptotic estimation of the entire state vector. Therefore, it is assumed that the dynamic system associated to Σ is strongly detectable. It implies, from Theorem 3.6 and Lemma 3.7.b, that $\operatorname{rank}\begin{bmatrix} D_1^T & F^T \end{bmatrix}^T = q$. Hence, from (3.8), $\bar{w}$ can be rewritten as $\bar{w} = \begin{bmatrix} D_1 \\ F \end{bmatrix}^+ \begin{bmatrix} \dot{\bar{x}}_1 - A_1\bar{x}_1 \\ y - C_1\bar{x}_1 \end{bmatrix}$, and its substitution into $\dot{\bar{x}}_2$ gives

$$\dot{\bar{x}}_2 = A_4\bar{x}_2 + A_2\bar{x}_1 + D_2 \begin{bmatrix} D_1 \\ F \end{bmatrix}^+ \begin{bmatrix} \dot{\bar{x}}_1 - A_1\bar{x}_1 \\ y - C_1\bar{x}_1 \end{bmatrix}.$$

Now, let $\hat{z}_2$ be the state observer for $\bar{x}_2$ defined by

$$\hat{z}_2 = \tilde{z}_2 + D_2 \begin{bmatrix} D_1 \\ F \end{bmatrix}^+ \begin{bmatrix} \bar{x}_1 \\ 0 \end{bmatrix}$$

$$\dot{\tilde{z}}_2 = A_4\hat{z}_2 + A_2\bar{x}_1 - D_2 \begin{bmatrix} D_1 \\ F \end{bmatrix}^+ \begin{bmatrix} A_1\bar{x}_1 \\ C_1\bar{x}_1 - y \end{bmatrix},$$

where $\bar{x}_1$ is supposed to be reconstructed from (3.11) using the recursive method given in 3.4.1. Thus, the error $\bar{x}_2 - \hat{z}_2$ is governed by the equation $\dot{\bar{x}}_2 - \dot{\hat{z}}_2 = A_4(\bar{x}_2 - \hat{z}_2)$. By the assumption that (3.1) is strongly detectable, from Theorem 3.6 and Lemma 3.7, we have

$$\hat{z}_2(t) \underset{t\to\infty}{\rightarrow} \bar{x}_2(t).$$

The previous scheme together with the definition of strong detectability gives rise to the following result.

REMARK 3. *Using the output of the system and a linear combinations of its derivatives, the strong detectability turns out to be a necessary and sufficient condition for the asymptotic estimation of x.*

4. Necessary and sufficient conditions for the estimation of $w(t)$.

In this section, we will show that for the estimation of w the system may be nonstrongly detectable, even further the pair (A, C) may be nondetectable. We will show that the necessary and sufficient condition to estimate w has to do with the set of invariant zeros of $\Sigma = (A, C, D, F)$ and the set of eigenvalues related to the unobservability of the pair (A, C). For that purpose we decompose the dynamics of the vector $\bar{x}_2$ in (3.8), where the second part of this decomposition corresponds to the unobservable part of (A, C). Since the proof of sufficiency of theorem establishing the conditions under which the estimation of w can be carried out is constructive, we give at the same time the main procedure for the estimation of w.

Let x_{w,x_0} be the solution of the differential equation $\dot{x}(t) = Ax(t) + Dw$, $x(0) := x_0$. Let $y_{w,x_0}(t) = Cx_{w,x_0} + Fw$. Thus, $\bar{x}_{w,x_0} = Px_{w,x_0}$ is governed by the set of equations (3.8). Now, let us recall the definition of left invertibility. The left invertibility concept in the time domain framework can be found in [6] for the case $F = 0$, and in [26] for the general class of inputs that are impulsive-smooth distributions. The definition given below is quite similar to the second one.

DEFINITION 4.1. *The system Σ is called **left invertible** if for any $w_1(t)$, $w_2(t) \in \mathbb{R}^q$ the following statement holds: $y_{w_1,x_0}(t) = y_{w_2,x_0}(t)$ for all $t \geq 0$ implies $w_1(t) = w_2(t)$ for all $t \geq 0$.*

It is clear that left invertibility is a necessary condition for the estimation of $w(t)$. However, we will see afterwards that it is not a sufficient one. That is quite obvious because the fulfilment of the left invertibility property depends on the knowledge of x_0, which is not the case considered here.

LEMMA 4.2. *Σ is left invertible if, and only if,* $\operatorname{rank}\left[\begin{smallmatrix}D_1\\F\end{smallmatrix}\right] = q$*.*

The next corollary follows directly from Lemmas 3.7 and 4.2.

COROLLARY 4.3. *Σ is left invertible if, and only if, the set of invariant zeros of Σ is finite.*

Let $\mathcal{N}^*$ be the unobservable subspace corresponding to the pair (A, C), that is, the greatest subspace satisfying

$$A\mathcal{N}^* \subset \mathcal{N}^* \text{ and } C\mathcal{N}^* = 0. \tag{4.1}$$

It is clear by the definition of $\mathcal{V}^*_\Sigma$ that $\mathcal{N}^* \subset \mathcal{V}^*_\Sigma$. Let $\mathcal{O}$ be the observability matrix of the pair (A, C); it is well known that $\mathcal{N}^* = \ker \mathcal{O}$.

Let N be a full column rank matrix whose columns form a basis of $\mathcal{N}^*$. Thus, we can chose a full column rank matrix V forming a basis of $\mathcal{V}^*$ adapted to $\mathcal{N}^*$, that is, V must have the following form

$$V = \left[\begin{array}{cc} \bar{V} & N \end{array}\right]. \tag{4.2}$$

Defining $n_2 := \dim \mathcal{N}^*$, we have that $\bar{V} \in \mathbb{R}^{n\times n-(n_1+n_2)}$, $N \in \mathbb{R}^{n\times n_2}$.

PROPOSITION 4.4. *If* $\operatorname{rank}\left[\begin{smallmatrix}D_1\\F\end{smallmatrix}\right] = q$ *and V has the form* (4.2), *the matrices K^* and Q satisfying* (3.6) *take the form*

$$K^* = \left[\begin{array}{cc} K_1^* & 0 \end{array}\right], \; Q = \left[\begin{array}{cc} Q_1 & 0 \\ Q_2 & Q_4 \end{array}\right] \tag{4.3}$$

for some matrices $K_1^* \in \mathbb{R}^{q\times(n-n_1-n_2)}$, $Q_1 \in \mathbb{R}^{(n-n_1-n_2)\times(n-n_1-n_2)}$, $Q_2 \in \mathbb{R}^{n_2\times(n-n_1-n_2)}$, and $Q_4 \in \mathbb{R}^{n_2\times n_2}$.

Thus, under the assumption that $\operatorname{rank}\left[\begin{smallmatrix}D_1\\F\end{smallmatrix}\right] = q$, and taking into account (3.7), (4.2), and (4.3), we have that the matrix A_4 in (3.8) takes the following partitioned form

$$A_4 := V^+\left(A + D\bar{K}^*\right)V = \left[\begin{array}{cc} Q_1 & 0 \\ Q_2 & Q_4 \end{array}\right] =: \left[\begin{array}{cc} A_{41} & 0 \\ A_{42} & A_{44} \end{array}\right], \tag{4.4}$$

where $A_{41} := Q_1$, $A_{42} := Q_2$, and $A_{44} := Q_4$. Therefore, partitioning the vector $\bar{x}_2 =: \left[\begin{smallmatrix}\bar{x}_{21}(t)\\ \bar{x}_{22}(t)\end{smallmatrix}\right]$ and from (4.3) and (4.4), the system (3.8) can be rewritten as

$$\begin{aligned} \begin{bmatrix} \dot{\bar{x}}_1(t) \\ \dot{\bar{x}}_{21}(t) \\ \dot{\bar{x}}_{22}(t) \end{bmatrix} &= \begin{bmatrix} A_1 & 0 & 0 \\ A_{21} & A_{41} & 0 \\ A_{22} & A_{42} & A_{44} \end{bmatrix} \begin{bmatrix} \bar{x}_1(t) \\ \bar{x}_{21}(t) \\ \bar{x}_{22}(t) \end{bmatrix} + \begin{bmatrix} D_1 \\ D_{21} \\ D_{22} \end{bmatrix} \bar{w}(t) \\ y(t) &= C_1\bar{x}_1(t) + F\bar{w}(t) \\ \bar{w}(t) &= w(t) - K_1^*\bar{x}_{21}(t), \end{aligned} \tag{4.5}$$

where $\bar{x}_{21} \in \mathbb{R}^{n-(n_1+n_2)}$ and $\bar{x}_{22} \in \mathbb{R}^{n_2}$. The matrices A_2 and D_2 given in (3.8) were partitioned as follows:

$$\begin{bmatrix} A_{21} \\ A_{22} \end{bmatrix} := A_2, \; \begin{bmatrix} D_{21} \\ D_{22} \end{bmatrix} := D_2.$$

First, we will show some facts that will be important in the procedure for finding the conditions under which we can estimate w.

Definition 4.5 (see [26]). *The constant $\lambda \in \mathbb{C}$ is said to be an (A, C)-unobservable eigenvalue if* $\operatorname{rank}\left[\begin{smallmatrix}\lambda I - A \\ C\end{smallmatrix}\right] < n$.

Lemma 4.6. *If* $\operatorname{rank}\left[\begin{smallmatrix}D_1 \\ F\end{smallmatrix}\right] = q$, *then:*

(a) *the set of (A, C)-unobservable eigenvalues is identical to the set of eigenvalues of A_{44}, and*

(b) *the set of invariant zeros of Σ that do not belong to the set of (A, C)-unobservable eigenvalues is identical to the set of eigenvalues of A_{41}.*

Theorem 4.7. *The following claims are equivalent.*

(i) *For any initial condition $x(0)$,*

$$y(t) = 0 \text{ for all } t \geq 0 \text{ implies } w(t) = 0 \text{ for all } t \geq 0. \tag{4.6}$$

(ii) *The set of invariant zeros of Σ is identical to the set of (A, C)-unobservable eigenvalues.*

(iii) *Σ is left invertible and $\mathcal{V}^* \equiv \mathcal{N}^*$.*

(iv)[5] $\operatorname{rank}\left[\begin{smallmatrix}D_1 \\ F\end{smallmatrix}\right] = q$ *and* $\operatorname{rank} M_{n,\Sigma} = \operatorname{rank} \mathcal{O}$.

Theorem 4.8. *The following sentences are equivalent.*

(i) *For any initial condition $x(0)$,*

$$y(t) = 0 \text{ for all } t \geq 0 \text{ implies } w(t) \underset{t\to\infty}{\to} 0. \tag{4.7}$$

(ii) *The set of invariant zeros of Σ that do not belong to the set of (A, C)-unobservable eigenvalues is in $\mathbb{C}^-$.*

(iii) $\operatorname{rank}\left[\begin{smallmatrix}D_1 \\ F\end{smallmatrix}\right] = q$ *and the set of eigenvalues of A_{41} is in $\mathbb{C}^-$.*

The following theorems establish the conditions, in terms of the invariant zeros of $\Sigma := (A, C, D, F)$ and the (A, C)-unobservable eigenvalues, under which the estimation of $w(t)$ can be carried out.

Theorem 4.9. *Based on the measurement of $y(t)$, the vector w can be estimated if, and only if, the set of invariant zeros of Σ that do not belong to the set of (A, C)-unobservable eigenvalues is in $\mathbb{C}^-$.*

Theorem 4.10. *Based on the measurement of $y(t)$, the vector w can be reconstructed in finite time if, and only if, the set of invariant zeros of Σ is identical to the set of (A, C)-unobservable eigenvalues.*

We should notice that if, in addition to the condition of Theorem 4.9, the system Σ satisfies the condition $\operatorname{rank}\left[\begin{smallmatrix}CD & F \\ F & 0\end{smallmatrix}\right] = \operatorname{rank} F + q$, then one can avoid using derivatives for the estimation of w. Because of, in such a case, $\bar{x}_1$ can be estimated asymptotically by using a linear observer (see, e.g., [16]). This can be summarized in the following theorem.

Theorem 4.11. *The vector w can be estimated from the system output y, without using any derivatives, if, and only if, the following two conditions are fulfilled:*

(1) *the set of invariant zeros of Σ that do not belong to the set of (A, C)-unobservable eigenvalues is in $\mathbb{C}^-$, and*

(2) $\operatorname{rank}\left[\begin{smallmatrix}CD & F \\ F & 0\end{smallmatrix}\right] = \operatorname{rank}(F) + q$.

5. Design of a robust observer.

The following restriction will be assumed to be satisfied throughout this section.

A1 The dynamic system associated to $\Sigma = (A, C, D, F)$ is strongly detectable.

Now, we will apply the scheme of design proposed in section 3 to the system (2.1) for the state vector estimate.

[5] M_n is given by (3.3) and $\mathcal{O}$ is the observability matrix of (A, C).

Thus, with P selected according to (3.5), after defining $\bar{x} := Px$, and with the partition $\bar{x} =: \left[\begin{smallmatrix}\bar{x}_1\\ \bar{x}_2\end{smallmatrix}\right]$, we have

$$
\begin{array}{c}
\begin{bmatrix} \dot{\bar{x}}_1(t) \\ \dot{\bar{x}}_2(t) \end{bmatrix} = \begin{bmatrix} A_1 & 0 \\ A_2 & A_4 \end{bmatrix} \begin{bmatrix} \bar{x}_1(t) \\ \bar{x}_2(t) \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} u + \begin{bmatrix} D_1 \\ D_2 \end{bmatrix} \bar{w}(t) \\
y(t) = C_1 \bar{x}_1(t) + F\bar{w}(t) \\
\bar{w}(t) = w(t) - K^* \bar{x}_2(t),
\end{array}
\tag{5.1}
$$

where the system and distribution matrices are defined according to (3.9), and $B_1 = M_{n,\Sigma}B$ and $B_2 = V^+ B$.

5.1. Bounding term. In the recursive method given in 3.4.1, some derivatives on time are needed; here, we suggest to use the super-twisting algorithm for the obtaining of the required derivatives. However, the super-twisting algorithm requires some bound of the state vector that is to be reconstructed; therefore, for ensuring the bound required we will use the following Luenberger observer.

$$
\dot{z} = PAP^{-1}z + PBu + L\left(y - CP^{-1}z\right)
\tag{5.2}
$$

The matrix $PAP^{-1} - LCP^{-1}$ must be Hurwitz. Such a requirement can always be satisfied, and A1 and the Proposition 3.8 guarantee its fulfilling. Defining $\bar{e} = \bar{x} - z$, we get the inequality $\|\bar{e}\| \leq \gamma \exp(-\lambda t) \|\bar{e}(0)\| + \mu w^+$ for some positive constants γ, λ, and μ. Now, let us make a partition of $\bar{e}$ into two vectors, i.e., $\bar{e}_1 = \bar{x}_1 - z_1$ and $\bar{e}_2 = \bar{x}_2 - z_2$, where $z^T =: \begin{bmatrix} z_1^T & z_2^T \end{bmatrix}$ and $z_1 \in \mathbb{R}^{n_1}$, $z_2 \in \mathbb{R}^{n-n_1}$. Let ζ be a constant satisfying $\zeta > \mu w^+$, then, after a finite time T, $\bar{e}_1$ and $\bar{e}_2$ stay bounded, i.e., $\|\bar{e}_1(t)\| < \zeta$ and $\|\bar{e}_2(t)\| < \zeta$, for all $t \geq T$.

5.2. Reconstruction of $M_{n_1,\bar{\Sigma}}\bar{e}_1(t)$. Now, the state estimation procedure of 3.4.1 will be applied to $\bar{e}_1$. Thus, once $\bar{e}_1$ is reconstructed, $\bar{x}_1$ can be recovered by the formula $\bar{x}_1 = \bar{e}_1 + z_1$. Firstly, recall that $\bar{\Sigma} := (A_1, C_1, D_1, F)$. Now, let us design the auxiliary vector σ defined by the equation

$$
\dot{\sigma}(t) = A_1 z_1(t) + B_1 u.
\tag{5.3}
$$

Define the first sliding variable s^1 as follows:

$$
s^1(t) = \bar{M}_{2,\bar{\Sigma}}^{\perp\perp} T_{1,\bar{\Sigma}} \begin{bmatrix} \left(F^{\perp}C_1\right)^{\perp\perp} F^{\perp}\left(y(t) - C_1\sigma(t)\right) \\ \int_0^t \left(y(\tau) - C_1 z(\tau)\right) d\tau \end{bmatrix} - \int_0^t v^1(\tau)\, d\tau.
$$

Thus, taking the derivative of s on time, and because of (3.3), we have

$$
\dot{s}^1(t) = M_{2,\bar{\Sigma}}\bar{e}_1(t) - v^1(t).
\tag{5.4}
$$

We design the output injection vector v^1 using the super-twisting technique ([17], [18]), involving not only a sign function but also its integral, that is,

$$
\begin{array}{l}
v_i^1 = \bar{v}_i^1 + \lambda_1 \left|s_i^1\right|^{1/2} \operatorname{sign} s_i^1 \\
\dot{\bar{v}}_i^1 = \alpha_1 \operatorname{sign} s_i^1,
\end{array}
\tag{5.5}
$$

where v_i^1 is the ith term of the vector v^1, and the same applies for $\bar{v}_i^1$ and s_i^1. The constants α_1 and λ_1 are selected to satisfy the inequalities:

$$
\varkappa_1 \geq \left\|M_{2,\bar{\Sigma}}\right\| \left(\left\|PAP^{-1} - LCP^{-1}\right\| \zeta + \|PD - LF\| w^+\right)
$$

$$
\alpha_1 > \varkappa_1,\ \lambda_1 > \frac{(1+\theta)(\varkappa_1 + \alpha_1)}{1-\theta} \sqrt{\frac{2}{\varkappa_1 - \alpha_1}},\ 1 > \theta > 0,
$$

where ζ was defined in subsection 5.1. Thus, according to [18], we have a second-order sliding mode, that is, $s^1(t) = \dot{s}^1(t) = 0$ for all $t \geq t_1$ where t_1 is the reaching time to the sliding mode. Therefore, from (5.4) and (5.5), we have that

$$\bar{v}^1(t) = M_{2,\bar{\Sigma}}\bar{e}_1(t), \text{ for all } t \geq t_1. \tag{5.6}$$

We can follow a quite similar scheme for the reconstruction of $M_{3,\bar{\Sigma}}\bar{e}_1(t)$. Namely, design the variable $s^2(t)$ as

$$s^2(t) = \bar{M}^{\perp\perp}_{3,\bar{\Sigma}} T_{2,\bar{\Sigma}} \begin{bmatrix} \bar{v}^1(t) - M_{2,\bar{\Sigma}}(\sigma(t) - z_1(t)) \\ \int_0^t (y(\tau) - C_1 z(\tau))\, d\tau \end{bmatrix} - \int_0^t v^2(\tau)\, d\tau.$$

Hence, taking into account (3.3) and (5.6), for $t \geq t_1$, the derivative of $s^2(t)$ is

$$\dot{s}^2(t) = M_{3,\bar{\Sigma}}\bar{e}_1(t) - v^2(t). \tag{5.7}$$

Again, the output injection vector v^2 is designed using the super-twisting algorithm,

$$\begin{aligned} v_i^2 &= \bar{v}_i^2 + \lambda_2 \left|s_i^2\right|^{1/2} \operatorname{sign} s_i^2 \\ \dot{\bar{v}}_i^2 &= \alpha_2 \operatorname{sign} s_i^2. \end{aligned} \tag{5.8}$$

The positive constants α_2 and λ_2 should satisfy the following upper bounds:

$$\varkappa_2 \geq \left\|M_{3,\bar{\Sigma}}\right\| \left(\left\|PAP^{-1} - LCP^{-1}\right\| \zeta + \|PD - LF\| w^+\right)$$
$$\alpha_2 > \varkappa_2,\ \lambda_2 > \tfrac{(1+\theta)(\varkappa_2+\alpha_2)}{1-\theta}\sqrt{\tfrac{2}{\varkappa_2-\alpha_2}},\ 1 > \theta > 0.$$

Then, according to [18], we have that $s^2(t) = \dot{s}^2(t) = 0$ for all t after t_2, which is the reaching time to the second sliding mode. Hence, in view of (5.7) and (5.8), we achieve the equality $\bar{v}^2(t) = M_{3,\bar{\Sigma}}\bar{e}_1(t)$ for all $t \geq t_2$.

We can generalize the previous procedure for the reconstruction of $M_{k,\bar{\Sigma}}\bar{e}_1(t)$ ($k = 2, \ldots, n_1 - 1$). Since, in this procedure, the main goal is the reconstruction of $\bar{e}_1(t)$, in the last step we will reconstruct directly $\bar{e}_1(t)$ instead of recovering $M_{n_1,\bar{\Sigma}}\bar{e}_1(t)$. The procedure is detailed below.

(a) Sliding variable s^1:

$$s^1(t) = \bar{M}^{\perp\perp}_{2,\bar{\Sigma}} T_{1,\bar{\Sigma}} \begin{bmatrix} \left(F^{\perp}C_1\right)^{\perp\perp} F^{\perp}(y(t) - C_1\sigma(t)) \\ \int_0^t (y(\tau) - C_1 z_1(\tau))\, d\tau \end{bmatrix} - \int_0^t v^1(\tau)\, d\tau; \tag{5.9}$$

sliding variable s^k, $k = 2, \ldots, n_1 - 2$:

$$s^k(t) = \bar{M}^{\perp\perp}_{k+1,\bar{\Sigma}} T_{k,\bar{\Sigma}} \begin{bmatrix} \bar{v}^{k-1}(t) - M_{k,\bar{\Sigma}}(\sigma(t) - z_1(t)) \\ \int_0^t (y(\tau) - C_1 z_1(\tau))\, d\tau \end{bmatrix} - \int_0^t v^k(\tau)\, d\tau; \tag{5.10}$$

sliding variable s^{n_1-1}:

$$\begin{aligned} s^{n_1-1}(t) = {} & \left[M_{n_1,\bar{\Sigma}}\right]^{-1} \bar{M}^{\perp\perp}_{n_1,\bar{\Sigma}} T_{n_1-1,\bar{\Sigma}} \begin{bmatrix} \bar{v}^{n_1-2}(t) - M_{n_1-1,\bar{\Sigma}}(\sigma(t) - z_1(t)) \\ \int_0^t (y(\tau) - C_1 z_1(\tau))\, d\tau \end{bmatrix} \\ & - \int_0^t v^{n_1-1}(\tau)\, d\tau, \end{aligned} \tag{5.11}$$

where $\sigma(t)$ is defined from (5.3).

(b) Output injection vector v^k $(k = 1, \ldots, n_1 - 1)$,

$$\begin{aligned} v_i^k &= \bar{v}_i^k + \lambda_k \left| s_i^k \right|^{1/2} \operatorname{sign} s_i^k \\ \dot{\bar{v}}_i^k &= \alpha_k \operatorname{sign} s_i^k \end{aligned} \tag{5.12}$$

being v_i^k the ith term of the vector v^k and $\bar{v}_i^k$ the ith term of the vector $\bar{v}^k$. The constants α_k and λ_k are designed according with [17] and [18]:

$$\begin{aligned} \varkappa_k &\geq \left\| M_{k+1,\bar{\Sigma}} \right\| \left(\left\| PAP^{-1} - LCP^{-1} \right\| \zeta + \left\| PD - LF \right\| w^+ \right), \quad k = 1, \ldots, n_1 - 2 \\ \varkappa_k &\geq \left\| PAP^{-1} - LCP^{-1} \right\| \zeta + \left\| PD - LF \right\| w^+, \qquad k = n_1 - 1 \\ \alpha_k &> \varkappa_k,\ \lambda_k > \frac{(1+\theta)(\varkappa_k + \alpha_k)}{1-\theta} \sqrt{\frac{2}{\varkappa_k - \alpha_k}},\ 1 > \theta > 0, \qquad k = 1, \ldots, n_1 - 1, \end{aligned}$$

where ζ was defined in 5.1 and w^+ is the bound of w. The procedure for the reconstruction of $\bar{e}_1(t)$ is given in the following theorem.

THEOREM 5.1 ([7]). *Following the design of s^k and v^k as in* (5.9)–(5.12), *we obtain the equalities*

$$\bar{v}^k(t) = M_{k+1,\bar{\Sigma}} \bar{e}_1(t) \text{ for all } t \geq t_k,\ k = 1, \ldots, n_1 - 2 \tag{5.13}$$

$$\bar{v}^{n_1 - 1}(t) = \bar{e}_1(t) \text{ for all } t \geq t_{n_1 - 1}, \tag{5.14}$$

where t_k is the reaching time to the kth sliding mode.

5.3. Observation of $\bar{x}_1$. Now, based on the recursive method given in 3.4.1, we have found the difference between the state vector and the Luenberger observer. It means that, following the method of design given previously in this section, we have that

$$\bar{x}_1(t) = z_1(t) + \bar{v}^{n_1 - 1}(t) \text{ for all } t \geq t_{n_1 - 1}. \tag{5.15}$$

The equality (5.15) motivates us to propose the reconstruction of the state $\bar{x}_1(t)$ by means of

$$\hat{z}_1(t) := z_1(t) + \bar{v}^{n_1 - 1}(t). \tag{5.16}$$

THEOREM 5.2. *Designing $\hat{z}_1(t)$ according to* (5.16), *we achieve the identity*

$$\hat{z}_1(t) = \bar{x}_1(t) \quad \text{for all } t \geq t_{n_1 - 1}. \tag{5.17}$$

Proof. It follows immediately by comparing (5.15) and (5.16). □

5.4. Observation of $\bar{x}_2$. Now, let us design an observer for the vector $\bar{x}_2$ given by (5.1). This is made by means of $\hat{z}_2$ which is designed as

$$\hat{z}_2 = \tilde{z}_2 + D_2 \begin{bmatrix} D_1 \\ F \end{bmatrix}^+ \begin{bmatrix} \hat{z}_1 \\ 0 \end{bmatrix}, \tag{5.18a}$$

$$\dot{\tilde{z}}_2 = A_4 \hat{z}_2 + A_2 \hat{z}_1 + B_2 u - D_2 \begin{bmatrix} D_1 \\ F \end{bmatrix}^+ \begin{bmatrix} A_1 \hat{z}_1 + B_1 u \\ C_1 \hat{z}_1 - y \end{bmatrix}. \tag{5.18b}$$

Thus, taking into account (5.17) we can obtain the dynamic equation for the error between $\bar{x}_2 - \hat{z}_2$, i.e.,

$$\dot{\bar{x}}_2(t) - \dot{\hat{z}}_2(t) = A_4 \left(\bar{x}_2(t) - \hat{z}_2(t) \right) \text{ for all } t \geq t_{n_1 - 1}.$$

Due to the Assumption A1, Theorem 3.6, and Lemma 3.7, the matrix A_4 is Hurwitz; therefore, the asymptotic stability of $\bar{x}_2 - \hat{z}_2$ is ensured, which implies

$$\hat{z}_2(t) \underset{t\to\infty}{\to} \bar{x}_2(t). \tag{5.19}$$

5.5. Observer for the original system. Thus, defining $\hat{z}^T = \begin{bmatrix} \hat{z}_1^T & \hat{z}_2^T \end{bmatrix}$, and from (5.16) and (5.19), we conclude that

$$\hat{z}(t) \underset{t\to\infty}{\to} \bar{x}(t). \tag{5.20}$$

Due to the coordinates change $\bar{x} = Px$ that we have used previously (P was defined in (3.5)), we have that the observer $\hat{x}$ for the original state vector has to be designed as

$$\hat{x}(t) = P^{-1}\hat{z}(t) = P^{-1}\begin{bmatrix} \hat{z}_1(t) \\ \hat{z}_2(t) \end{bmatrix} \tag{5.21}$$

with $\hat{z}_1$ and $\hat{z}_2$ defined from (5.16) and (5.18), respectively.

THEOREM 5.3. *The observer $\hat{x}$ given by* (5.21) *converges to the original state vector x. That is,*

$$\hat{x} \underset{t\to\infty}{\to} x(t).$$

Proof. It is clear from (5.20) and (5.21). □

6. Identification of unknown inputs $w(t)$ (General case). Consider again the system (2.1). Here, we apply the results obtained in section 4 for the estimation of the unknown inputs in the general case. That is, the proposed algorithm is not required to estimate the entire state vector since it is based on the necessary and sufficient conditions obtained at the end of section 4.

Using the transformation P defined according to (3.5), but with V selected according to (4.2), we have that the dynamic equations for the transformed system $\bar{x} = Px$ takes the form

(6.1)

$$\begin{bmatrix} \dot{\bar{x}}_1(t) \\ \dot{\bar{x}}_{21}(t) \\ \dot{\bar{x}}_{22}(t) \end{bmatrix} = \begin{bmatrix} A_1 & 0 & 0 \\ A_{21} & A_{41} & 0 \\ A_{22} & A_{42} & A_{44} \end{bmatrix} \begin{bmatrix} \bar{x}_1(t) \\ \bar{x}_{21}(t) \\ \bar{x}_{22}(t) \end{bmatrix} + \begin{bmatrix} B_1 \\ B_{21} \\ B_{22} \end{bmatrix} u + \begin{bmatrix} D_1 \\ D_{21} \\ D_{22} \end{bmatrix} \bar{w}(t)$$
$$y(t) = C_1\bar{x}_1(t) + F\bar{w}(t)$$
$$\bar{w}(t) = w(t) - K_1^*\bar{x}_{21}(t),$$

where $\bar{x}_1 \in \mathbb{R}^{n_1}$, $\bar{x}_{21} \in \mathbb{R}^{n-n_1-n_2}$, and $\bar{x}_{22?} \in \mathbb{R}^{n_2}$. The partitions of the system and distribution matrices comes from (3.9) and (4.4). The matrices not defined yet are $B_1 := M_{n,\Sigma}B$, $\left[\begin{smallmatrix} B_{21} \\ B_{22} \end{smallmatrix}\right] := V^+B$, $\left[\begin{smallmatrix} D_{21} \\ D_{22} \end{smallmatrix}\right] := V^+D$. Since in the section 4 we have found the conditions under which we can estimate $w(\cdot)$, throughout this section we will assume that:[6]

B1 The set of the invariant zeros of $\Sigma = (A, C, D, F)$ that do not belong to the set of the (A, C)-unobservable eigenvalues is in $\mathbb{C}^-$.

Moreover, the use of the super-twisting algorithm as a differentiator imposes other restrictions related to the smoothness and boundedness of $w(t)$, i.e.,

B2 There is a known constant α_w such that $\|\dot{w}(t)\| \le \alpha_w$.

[6] It should be noticed that the assumption B1 is a structural assumption; meanwhile B2 is an assumption required by the algorithm used (super-twisting) to estimate the needed derivatives.

Step 1.a Estimation of $\bar{x}_1$.

As was established in section 4, for the estimation of $w\left(t\right)$ it is enough to estimate the states $\bar{x}_1$ and $\bar{x}_{21}$ (even in the case when $\bar{x}_{22}$ cannot be estimated). Therefore, we can estimate the reduced vector $\begin{bmatrix}\bar{x}_1^T & \bar{x}_{21}^T\end{bmatrix}^T$ following the same procedure given in the previous section for estimating $\bar{x}$. In other words, to estimate $\bar{x}_1$ and $\bar{x}_{21}$, we should follow the procedure of the previous section, but using the reduced vector $\begin{bmatrix}\bar{x}_1^T & \bar{x}_{21}^T\end{bmatrix}^T$ instead of all the vector $\bar{x} = \begin{bmatrix}\bar{x}_1^T & \bar{x}_2^T\end{bmatrix}^T$. Thus, z in subsection 5.1 becomes $z := \left[\begin{smallmatrix}z_1\\ z_{21}\end{smallmatrix}\right] (z_1 \in \mathbb{R}^{n_1}$, $z_{21} \in \mathbb{R}^{n-n_1-n_2})$, and its dynamics is governed by the equations

$$\dot{z} = \bar{A}z - \bar{B}u + L\left(y - \bar{C}z\right), \tag{6.2}$$

where

$$\bar{A} = \begin{bmatrix} A_1 & -D_1K_1^* \\ A_{21} & A_{41} - D_{21}K_1^* \end{bmatrix}, \bar{B} = \begin{bmatrix} B_1 \\ B_{21} \end{bmatrix}, \bar{D} = \begin{bmatrix} D_1 \\ D_{21} \end{bmatrix}, \bar{C} = \begin{bmatrix} C_1 & -FK_1^* \end{bmatrix}.$$

Notice that

$$H\left(s\right) = \begin{bmatrix} sI - A_1 & -D_1K_1^* \\ A_{21} & sI - \left(A_{41} - D_{21}K_1^*\right) \\ C_1 & -FK_1^* \end{bmatrix} = \begin{bmatrix} sI - A_1 & 0 & D_1 \\ A_{21} & sI - A_{41} & D_{21} \\ C_1 & 0 & F \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & -K_1^* & I \end{bmatrix}$$

Since $\bar{\Sigma}$ has no invariant zeros (Lemma 3.4) and rank $\begin{bmatrix}D_1^T & F^T\end{bmatrix}^T = q$ (B1 and Theorem 4.8), the matrix $H\left(s\right)$ loses rank only for s being an eigenvalue of A_{41}. Thus, by the assumption B1 and Theorem 4.8, A_{41} is Hurwitz and, consequently, the pair $(\bar{A}, \bar{C})$ is detectable. Then, selecting the matrix $L \in \mathbb{R}^{(n-n_2)\times p}$ in such a way that $(\bar{A} - L\bar{C})$ is Hurwitz, we have, for $\bar{e}_1 := \bar{x}_1 - z_1$, the inequality $\left\|\bar{e}_1\right\| \leq \gamma \exp\left(-\lambda t\right) \left\|\bar{e}_1\left(0\right)\right\| + \mu w^+$ for some constants γ, λ, μ. Therefore, for ζ satisfying $\zeta > \mu w^+$, we obtain the inequality $\left\|\bar{e}_1\left(t\right)\right\| < \zeta$. Thus, we estimate $\bar{x}_1$ by means of $\hat{z}_1$ given from (5.16) that is designed following the same procedure used in (5.9)–(5.12), but with z_1 from (6.2).

Step 1.b

Estimation of $\bar{x}_{21}$.

The vector $\bar{x}_{21}$ must be estimated by means of $\hat{z}_{21}$ that has to be designed in the following form:

$$\hat{z}_{21} = \tilde{z}_{21} + D_{21}\begin{bmatrix}D_1\\ F\end{bmatrix}^+ \begin{bmatrix}\dot{\hat{z}}_1\\ 0\end{bmatrix}$$

$$\dot{\tilde{z}}_{21} = A_{41}\hat{z}_{21} + A_{21}\hat{z}_1 + B_{21}u - D_{21}\begin{bmatrix}D_1\\ F\end{bmatrix}^+ \begin{bmatrix}A_1\hat{z}_1 + B_1u\\ C_1\hat{z}_1 - y\end{bmatrix}$$

Thus, from (6.1), we have that the dynamic equation for the difference $\left(\bar{x}_{21} - \hat{z}_{21}\right)$ is $\frac{d}{dt}\left(\bar{x}_{21}\left(t\right) - \hat{z}_{21}\left(t\right)\right) = A_{41}\left(\bar{x}_{21}\left(t\right) - \hat{z}_{21}\left(t\right)\right)$, but the Assumption B1 and Theorem 4.8 implies that A_{41} is Hurwitz. Therefore,

$$\hat{z}_{21}\left(t\right) \underset{t\to\infty}{\rightarrow} \bar{x}_{21}\left(t\right). \tag{6.3}$$

Step 2 Estimation of $w(t)$

Let us define $r := \operatorname{rank} F$. If $r < q$, define $G \in \mathbb{R}^{q\times q}$ as a nonsingular matrix so that

$$\begin{bmatrix} D_1 \\ F \end{bmatrix} G = \begin{bmatrix} D_{11} & D_{12} \\ 0 & F_2 \end{bmatrix}, \; F_2 \in \mathbb{R}^{p\times r}, \; \operatorname{rank} F_2 = r. \tag{6.4}$$

If $r = q$, $G := I_{q\times q}$. Now, let us make a partition of G^{-1} as

$$G^{-1} =: \begin{bmatrix} \bar{G}_1 \\ \bar{G}_2 \end{bmatrix}, \, \bar{G}_1 \in \mathbb{R}^{(q-r)\times q}, \, \bar{G}_2 \in \mathbb{R}^{r\times q}. \tag{6.5}$$

Thus, from (6.4) and (6.5), we have

$$\begin{aligned} y(t) &= C_1 \bar{x}_1(t) + FGG^{-1}\bar{w}(t) \\ &= C_1 \bar{x}_1(t) + F_2 \bar{G}_2 \bar{w}(t). \end{aligned}$$

Hence, premultiplying the last equation by F_2^+, we obtain a linear combination of the rows of $\bar{w}$, i.e.,

$$\bar{G}_2 \bar{w}(t) = F_2^+ y(t) - F_2^+ C_1 \bar{x}_1(t). \tag{6.6}$$

Therefore, $\bar{G}_2 w(t)$ can be written as

$$\bar{G}_2 w(t) = F_2^+ y(t) - F_2^+ C_1 \bar{x}_1(t) + \bar{G}_2 K_1^* \bar{x}_{21}(t). \tag{6.7}$$

Now, let z_{w} be the state vector of the auxiliary system characterized by the equation

$$\begin{aligned} \dot{z}_{\mathrm{w}}(t) = A_1 \hat{z}_1(t) &+ B_1 u + D_{11}\left(u_{\mathrm{w}}(t) - \bar{G}_1 K_1^* \hat{z}_{21}(t)\right) \\ &+ D_{12}\left(F_2^+ y(t) - F_2^+ C_1 \hat{z}_1(t)\right). \end{aligned} \tag{6.8}$$

Let us estimate $\bar{G}_1 \bar{w}(t)$ using a sliding mode technique, specifically, the super-twisting. We design the sliding variable ξ in the following way:

$$\xi(t) = D_{11}^+ \left[\hat{z}_1(t) - z_{\mathrm{w}}(t)\right]. \tag{6.9}$$

Thus, in view of the identity $D_1\bar{w} = D_1 GG^{-1}\bar{w} = D_{11}\bar{G}_1\bar{w} + D_{12}\bar{G}_2\bar{w}$, from (6.1), (6.6), and (6.8), we achieve the equality

$$\dot{\xi}(t) = \bar{G}_1 w(t) - u_{\mathrm{w}}(t) - \bar{G}_1 K_1^* \left(\bar{x}_{21}(t) - \hat{z}_{21}(t)\right)$$

for all $t \geq t_{n_1-1}$. Then, using

$$\begin{gathered} u_{\mathrm{w}}(t) = \bar{u}_{\mathrm{w}}(t) + \lambda |\xi|^{1/2} \operatorname{sign} \xi \\ \dot{\bar{u}}_{\mathrm{w}}(t) = \alpha \operatorname{sign} \xi(t) \\ \alpha \geq \alpha_w, \; \lambda > \tfrac{1-\theta(\varkappa+\alpha)}{1+\theta}\sqrt{\tfrac{2}{\varkappa-\alpha}}, \; \varkappa - \alpha > 0, \; 0 < \theta < 1, \end{gathered}$$

there is a reaching time t_w to the second-order sliding mode ($\xi(t) = \dot{\xi}(t) = 0$, for all $t \geq t_w > t_{n_1-1}$). Hence, we get

$$\bar{u}_{\mathrm{w}}(t) = \bar{G}_1 w(t) - \bar{G}_1 K_1^* \left(\bar{x}_{21}(t) - \hat{z}_{21}(t)\right). \tag{6.10}$$

Thus, from (6.3),

$$\bar{u}_{\mathrm{w}}(t) \underset{t\to\infty}{\to} \bar{G}_1 w(t).$$

Thus, the estimate of $w(t)$ is done by means of

$$\hat{w}(t) = G\begin{bmatrix} \bar{u}_{\mathrm{w}}(t) \\ F_2^{+}y(t) - F_2^{+}C_1\hat{z}_1(t) + \bar{G}_2K_1^{*}\hat{z}_{21}(t) \end{bmatrix}. \tag{6.11}$$

In view of (6.7), (6.10), and (6.5), we achieve the equality

$$\hat{w}(t) = w(t) - K_1^{*}(\bar{x}_{21}(t) - \hat{z}_{21}(t)) + G\begin{bmatrix} 0 \\ F_2^{+}C_1(\bar{x}_1(t) - \hat{z}_1(t)) \end{bmatrix}.$$

However, from (5.17) and (6.3), we conclude that $\hat{w}(t)$ converges asymptotically to $w(t)$, i.e.,

$$\hat{w}(t) \underset{t\to\infty}{\to} w(t). \tag{6.12}$$

Remark 4. *It should be noticed that, for the case when* B1 *is fulfilled with* $\operatorname{rank} M_n = \operatorname{rank} O$ *(Theorem* 4.7*),* $\bar{x}^T = \begin{bmatrix} \bar{x}_1 & \bar{x}_{22} \end{bmatrix}$*. Therefore, the limit in* (6.12) *becomes in the equality* $\hat{w}(t) = w(t)$*, for all* $t \geq t_w > t_{n_1-1}$.

7. Numerical examples.

Here we give two numerical examples. The first one is to show the scheme of design for the estimation of the state of a strongly detectable system. The second example shows the scheme of design for the estimation of the unknown inputs of a system which is not strongly detectable.

7.1. Example 1.

Consider the following academic example. Let a linear system be governed by the following equations:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \\ \dot{x}_5 \end{bmatrix} = \underbrace{\begin{bmatrix} -2.51 & 0.33 & 0.68 & 1.12 & -0.25 \\ 0.14 & -0.23 & -0.31 & 0.91 & 0.36 \\ 0.51 & -1.18 & 0.41 & 0.63 & -0.77 \\ 0.22 & 0.33 & 0.46 & 0.65 & -0.77 \\ 0.23 & 0.33 & 3.97 & 0.06 & 0.69 \end{bmatrix}}_{A} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} + \underbrace{\begin{bmatrix} 0.43 \\ 0 \\ 0.92 \\ 1.20 \\ -1.27 \end{bmatrix}}_{B} (u + w_1)$$

$$y = \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}}_{C} x + \underbrace{\begin{bmatrix} 1 \\ 0 \end{bmatrix}}_{\tilde{F}} w_2$$

Defining $w = \begin{bmatrix} w_1 & w_2 \end{bmatrix}^T$, $D = \begin{bmatrix} B & 0_{5\times1} \end{bmatrix}$, and $F = \begin{bmatrix} 0_{2\times1} & \tilde{F} \end{bmatrix}$, this linear system takes the form of (2.1). In the simulations was used a control given by $u = -K\hat{x} + 1.5\sin(2t)$, $K = \begin{bmatrix} 0.57 & 5.66 & 1.25 & 6.94 & 1.68 \end{bmatrix}$. The unknown inputs are $w_1 = 2\sin(2t) + 0.47$ and $w_2 = -\sin(2t) + 0.53$.

It can be verified that the set of the invariant zeros of Σ is $\{-1.84 + 0.48i, -1.84 - 0.48i, -3.27\}$. Therefore, the system Σ is not strongly observable, but, from Theorem 3.6, it is strongly detectable.

Construction of the hierarchical observer for $\bar{x}_1$. The matrices $M_{1,\bar{\Sigma}}$ and $M_{2,\bar{\Sigma}}$, computed following (3.3) for $\Sigma = \bar{\Sigma}$, take the form $M_{1,\bar{\Sigma}} = \begin{bmatrix} 0 & 1 \end{bmatrix}$ and $M_{2,\bar{\Sigma}} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$. As we can anticipate from Lemma 3.4 the matrix $M_{2,\bar{\Sigma}}$ is invertible. We

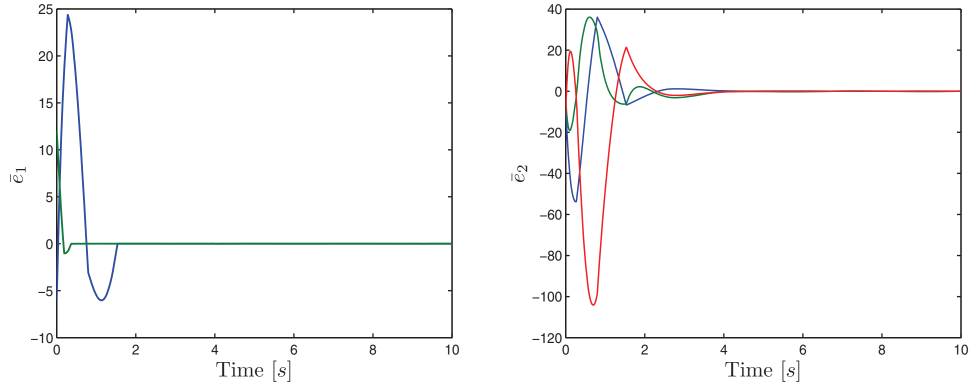


FIG. 7.1. *Error of observation* $\bar{e}_1 = \bar{x}_1 - \hat{z}_1$ *and* $\bar{e}_2 = \bar{x}_2 - \hat{z}_2$*, for Example* 1.

construct the Luenberger observer as in (5.2). Next, we construct σ as in (5.3). In this case $n_1 := \dim M_{2,\bar{\Sigma}} = 2$; therefore, it is needed to design only one sliding surface $s^1 \in \mathbb{R}^2$, which is designed as

$$s^1(t) = \underbrace{\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}}_{M_{2,\bar{\Sigma}}^{-1}} \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{\bar{M}_{2,\bar{\Sigma}}^{\perp\perp} T_{1,\bar{\Sigma}}} \begin{bmatrix} \overbrace{\begin{bmatrix} 0 & 1 \end{bmatrix}}^{\left(F^{\perp}C_1\right)^{\perp\perp}F^{\perp}} \quad (y(t) - C_1\sigma(t)) \\ \int_0^t (y(\tau) - C_1 z_1(\tau))\, d\tau \end{bmatrix} - \int_0^t v^1(\tau)\, d\tau$$

The matrices $\bar{M}_{2,\bar{\Sigma}}^{\perp\perp}$and $T_{1,\bar{\Sigma}}$ are designed following (3.3). The output injection v^1 takes the form

$$\begin{aligned} v_i^1 &= \bar{v}_i^1 + 20\left|s_i^1\right|^{1/2} \operatorname{sign} s_i^1 \\ \dot{\bar{v}}_i^1 &= 15 \operatorname{sign} s_i^1, \quad i = 1, 2 \end{aligned}$$

Thus, we have that the reconstruction of $\bar{x}_1$ is done by $\hat{z}_1(t) = z_1(t) + v_1^{(n_1-1)}(t)$.

The observer $\hat{z}_2(t)$ for $\bar{x}_2$ is designed according to (5.18), for this observer any gain is not needed to be calculated. In the Figure 7.1 the observation errors $\bar{e}_1 = \bar{x}_1 - \hat{z}_1$ and $\bar{e}_2 = \bar{x}_2 - \hat{z}_2$ are drawn. For the simulations we use a sampling step of 10^{-4}.

Then the hierarchical observer for the original state x is designed as $\hat{x} := P^{-1}\hat{z}$. The trajectories of $x(t)$ together with the trajectories of its observer $\hat{x}(t)$ are depicted in Figure 7.2.

7.2. Example 2. Consider the following nonstrongly detectable system.

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \\ \dot{x}_5 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix}}_{x} + \underbrace{\begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}}_{D} \underbrace{\begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}}_{w}$$

$$y = \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}}_{C} x + \underbrace{\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{F} w$$

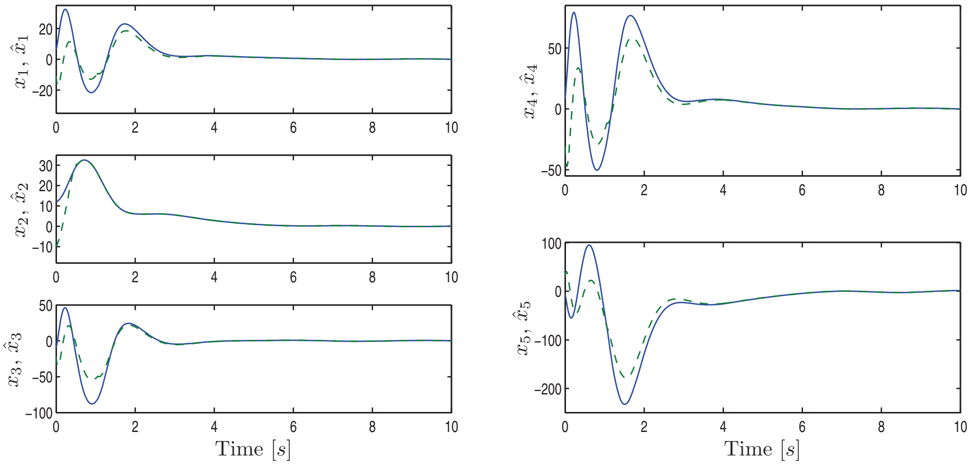


FIG. 7.2. *Trajectories of* $x(t)$ *(solid) and* $\hat{x}(t)$ *(dashed), for Example* 1.

Next we present the matrix V that forms a basis of $\mathcal{V}_\Sigma^*$ (see (3.4) and(4.2)) and the matrix N that forms a basis of $\mathcal{N}^*$ (see (4.1)). Also the matrix P that changes the coordinates of the system is written below (see (3.5)).

$$V = N = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \; P = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0.7071 & -0.7071 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Thus, by the change of coordinates $\bar{x} = Px$, we get the decomposition obtained in (4.5).

$$\begin{bmatrix} \dot{\bar{x}}_{1,1} \\ \dot{\bar{x}}_{1,2} \\ \dot{\bar{x}}_{1,3} \\ \dot{\bar{x}}_{1,4} \\ \dot{\bar{x}}_{22} \end{bmatrix} = \begin{bmatrix} 1 & -1.41 & 0 & 0 & 0 \\ 0.707 & -1 & -707 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \underset{A_{44}}{\boxed{0}} \end{bmatrix} \underbrace{\begin{bmatrix} \bar{x}_{1,1} \\ \bar{x}_{1,2} \\ \bar{x}_{1,3} \\ \bar{x}_{1,4} \\ \bar{x}_{22} \end{bmatrix}}_{\bar{x}} + \begin{bmatrix} 1 & 0 & 0 \\ 0.707 & -707 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}$$

$$y = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & -1.41 & 0 & 0 & 0 \end{bmatrix} \bar{x} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}$$

It can be verified that rank $\begin{bmatrix} D_1^T & F^T \end{bmatrix}^T = 3$, rank $M_{4,\Sigma} = \operatorname{rank} O = 4$. Thus, the condition of theorem 4.9 is accomplished, which implies that w can be reconstructed in a finite time. In this case the weakly unobservable subspace corresponding to $\Sigma = (A, C, D, F)$ and the unobservable subspace corresponding to (A, C) are identical. It means that $A_4 = A_{44}$ and, consequently, $\bar{x}_{21}$ does not exist. Therefore, for the reconstruction of the w only $\bar{x}_1$ has to be reconstructed.

Nevertheless, since $A_{44} = 0$, according to Lemma 3.7, the system Σ has only one invariant zero, which is equal to zero. Hence, Σ is nonstrongly detectable, but also notice that (A, C) is nondetectable. Hence, the state vector cannot be estimated neither in finite time nor asymptotically.

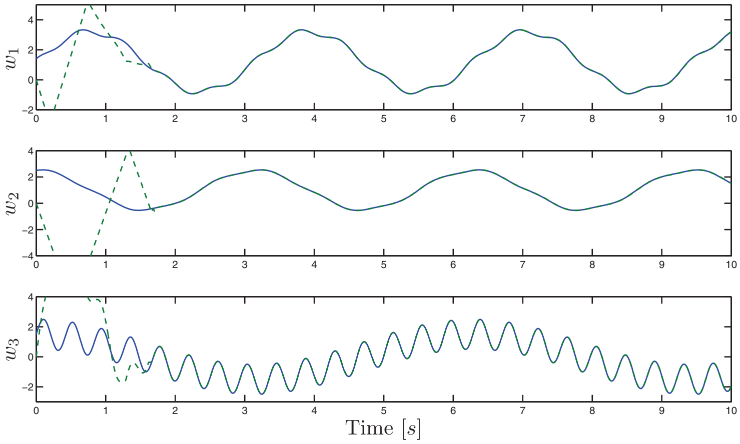


FIG. 7.3. *Comparison between w_i (solid) and their estimate $\hat{w}_i$ (dashed), $i = 1, 2, 3$, for Example* 2.

Using the method proposed in section 6 we can estimate the unknown inputs vector w. Firstly, for the estimation of $\bar{x}_1$ the use of two sliding surfaces was needed, s^1 and s^2, designed according to 5.9 and 5.11. The next step was the estimation of w. The estimate of w_i is given by $\hat{w}_i$ $(i = 1, 2, 3)$, respectively, and it is shown in Figure 7.3.

Conclusions. We have shown that, for a system with unknown inputs appearing explicitly in both the state equations and the system output, the *strong detectability* is a *necessary and sufficient condition* for the estimation of the original state vector. Since if the system is not strongly observable, it is impossible to design a standard differential observer providing state estimate. Hence, we have proposed another approach related to the design of an algebraic-type observer. We have shown that the suggested approach can successfully work for both strongly observable and strongly detectable systems. Thus, we have proposed to decompose the system into two subsystems. The first one is strongly observable for the zero control input. The second one is not strongly observable but can be detected. Thus, in the new coordinates, one uses the output of the system and its derivatives unaffected by the unknown input to reconstruct the state vector of the first subsystem. For the second subsystem one needs to design an observer that converges asymptotically to the state vector of the second subsystem. This scheme of design brings as a result an observer whose trajectories converge to those of the original state vector and whose rate of convergence does not depend on the unknown inputs.

Furthermore, we have shown that left invertibility is not a sufficient condition under which the estimation can be carried out. But also we have shown that a system can be nonstrongly detectable, even nondetectable as we saw in the Example 2, and the estimation of the unknown inputs can still be carried out.

Perhaps as the most important result of this paper, we have proven that the *necessary and sufficient condition* under which the estimation of the unknown inputs can be carried out is that the set of *the invariant zeros* of the system (with respect to the unknown inputs) that do not belong to the set of *unobservable eigenvalues* is in the interior of the set of complex numbers with negative real part. Based on these results we have proposed a scheme of design for the estimation of the unknown inputs.

Appendix. Proofs of propositions, lemmas, and theorems.

Proof of Lemma 3.4. From (3.10), $\dim\ker M_{n,\Sigma_{\bar{K}^*,P}} = \dim\ker M_{n,\Sigma}$. Then, applying (3.3) for calculating $M_{n,\Sigma_{\bar{K}^*,P}}$, we get that $M_{n,\Sigma_{\bar{K}^*,P}} = \begin{bmatrix} M_{n_1,\bar{\Sigma}} & 0 \end{bmatrix}$, where $M_{n,\Sigma_{\bar{K}^*,P}} \in \mathbb{R}^{n_1\times n}$, $M_{n_1,\bar{\Sigma}} \in \mathbb{R}^{n_1\times n_1}$. Taking into account that $\operatorname{rank} M_{n,\Sigma_{\bar{K}^*,P}} = n_1$, one can conclude that $\ker M_{n_1,\bar{\Sigma}} = \mathcal{V}^*_{\bar{\Sigma}} = 0$. □

Proof of Lemma 3.7. From (3.8) and by a rearranging of matrices we get

$$\operatorname{rank} P(s) = \operatorname{rank} \begin{bmatrix} P & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} sI - A & -D \\ C & F \end{bmatrix} \begin{bmatrix} P^{-1} & 0 \\ \bar{K}^* P^{-1} & I \end{bmatrix}$$
$$= \operatorname{rank} \begin{bmatrix} sI - P\left(A + D\bar{K}^*\right)P^{-1} & -PD \\ \left(C + F\bar{K}^*\right)P^{-1} & F \end{bmatrix} = \operatorname{rank} \begin{bmatrix} sI - A_1 & -D_1 & 0 \\ C_1 & F & 0 \\ -A_2 & -D_2 & sI - A_4 \end{bmatrix}.$$

From Lemma 3.4 and Fact 1, $\bar{\Sigma} := (A_1, C_1, D_1, F)$ has no invariant zeros. This means that for the case $\operatorname{rank} \begin{bmatrix} D_1^T & F^T \end{bmatrix}^T = q$ the only way that the previous arrangement of matrices can lose rank is when s is an eigenvalue of A_4. This proves the clause a).

On the other hand, if $\operatorname{rank} \begin{bmatrix} D_1^T & F^T \end{bmatrix}^T < q$, there exists a nonsingular matrix[7] $G \in \mathbb{R}^{q\times q}$ so that $\begin{bmatrix} -D_1 \\ F \end{bmatrix} G = \begin{bmatrix} H_1 & 0 \\ H_2 & 0 \end{bmatrix}$. Hence,

$$\operatorname{rank} \bar{P}(s) = \operatorname{rank} \begin{bmatrix} sI - A_1 & H_1 & 0 & 0 \\ C_1 & H_2 & 0 & 0 \\ -A_2 & D_{21} & D_{22} & sI - A_4 \end{bmatrix}$$

with $\bar{P}(s)$ defined in (3.2) and $\begin{bmatrix} D_{21} & D_{22} \end{bmatrix} := -D_2 G$. Thus, $\bar{P}(s)$ loses rank for every $s \in \mathbb{C}$, so the clause b) is proven. □

Proof of Lemma 4.2. *Necessity*: suppose $\operatorname{rank} \begin{bmatrix} D_1^T & F^T \end{bmatrix}^T < q$. Then there is a constant vector $v \in \mathbb{R}^q$, $v \neq 0$, so that $\begin{bmatrix} D_1^T & F^T \end{bmatrix}^T v = 0$. Let us choose $w_1(t) = K^*\bar{x}_2(t) + v$ and $w_2(t) = K^*\bar{x}_2(t)$. Thus, for $x_0 = 0$, from (3.8), we have that $y_{w_1,x_0}(t) = y_{w_2,x_0}(t) = 0$; meanwhile, $w_1(t) - w_2(t) = v \neq 0$. Thus, the necessity is proven.

Sufficiency: suppose $\operatorname{rank} \begin{bmatrix} D_1^T & F^T \end{bmatrix}^T = q$. Let $w_1(t)$ and $w_2(t)$ be two inputs so that $y_{w_1,x_0}(t) = y_{w_2,x_0}(t)$ for all $t \geq 0$. Now, notice that the equality $x_{w_1,x_0}(t) - x_{w_2,x_0}(t) = x_{0,w_1-w_2}(t)$ is valid for any initial condition x_0 (by notation, $x_{0,w_1-w_2}(0) = 0$). Thus, the initial condition for the transformed system $\bar{x}_{0,w_1-w_2}(t)$ with unknown input $w_1(t) - w_2(t)$ is $\bar{x}_{0,w_1-w_2}(0) = Px_{0,w_1-w_2} = 0$. Furthermore, we have that $y_{0,w_1-w_2}(t) = y_{w_1,x_0}(t) - y_{w_2,x_0}(t) = 0$ for all $t \geq 0$. This, due to the fact that $\bar{\Sigma} = (A_1, C_1, D_1, F)$ is strongly observable, implies that $\bar{x}_{1_{0,w_1-w_2}}(t) = 0$ for all $t \geq 0$, which, from (3.8), leads to the equality

$$w_1(t) - w_2(t) - K^*\bar{x}_{2,w_1-w_2}(t) = 0, \text{ for all } t \geq 0 \tag{A.1}$$

Therefore, we have that $\dot{\bar{x}}_{2,w_1-w_2}(t) = A_4\bar{x}_{2,w_1-w_2}(t)$, and since $\bar{x}_{w_1-w_2}(0) = 0$, we get $\bar{x}_{2,w_1-w_2}(t) \equiv 0$. The last equality and (A.1) imply that $w_1(t) = w_2(t)$, which proves the sufficiency. □

[7] Actually, G is used to divide the matrix $J := \begin{bmatrix} -D_1 \\ F \end{bmatrix}$. Indeed, let G_2 be a matrix whose columns span the kernel of J, and let G_1 be a matrix so that $G = \begin{bmatrix} G_1 & G_2 \end{bmatrix}$ is not singular. Thus, $JG = \begin{bmatrix} H & 0 \end{bmatrix}$.

Proof of Proposition 4.4. Partitioning K^* and Q as $K^* =: \left[\begin{array}{cc} K_1^* & K_2^* \end{array}\right]$ and $Q =: \left[\begin{smallmatrix} Q_1 & Q_3 \\ Q_2 & Q_4 \end{smallmatrix}\right]$, the equations in (3.6) can be rewritten in the form

$$\begin{aligned} A\left[\begin{array}{cc} \bar{V} & N \end{array}\right] + D\left[\begin{array}{cc} K_1^* & K_2^* \end{array}\right] &= \left[\begin{array}{cc} \bar{V}Q_1 + NQ_2 & \bar{V}Q_3 + NQ_4 \end{array}\right] \\ C\left[\begin{array}{cc} \bar{V} & N \end{array}\right] + F\left[\begin{array}{cc} K_1^* & K_2^* \end{array}\right] &= 0 \end{aligned}$$

From there we can obtain the equations

$$AN + DK_2^* = \bar{V}Q_3 + NQ_4 \tag{A.2}$$

$$CN + FK_2^* = 0 \tag{A.3}$$

Taking into account that $M_{n,\Sigma}V = 0$ and $CN = 0$, we achieve the identities

$$\begin{aligned} M_{n,\Sigma}AN + D_1K_2^* &= 0 \\ FK_2^* &= 0 \end{aligned}$$

Furthermore, since N spans $\mathcal{N}^*$, $A\mathcal{N}^* \subset \mathcal{N}^* \subset \mathcal{V}_\Sigma^*$, and $M_{n,\Sigma}\mathcal{V}_\Sigma^* = 0$, then, $M_{n,\Sigma}AN = 0$. Therefore, $\left[\begin{smallmatrix} D_1 \\ F \end{smallmatrix}\right]K_2^* = 0$, which, from the first assumption of the proposition, implies $K_2^* = 0$. Moreover, since the span of AN belongs to the span of N and because of $\bar{V}$ and N are linearly independent, from (A.2) we conclude that $Q_3 = 0$. □

Proof of Lemma 4.6. For V given by (4.2) and from the Proposition 4.4, we have $\bar{K}^*N = K^*V^+N = 0$. Furthermore, $\left(A + D\bar{K}^*\right)V = VQ$; thus, from (4.3) and (4.4), we get $AN = NA_{44}$. Hence, with $P^{-1} = \left[\begin{array}{ccc} M_{n,\Sigma}^+ & \bar{V} & N \end{array}\right]$, we can decompose the pair (A, C) in its observable and unobservable part. Indeed, first let us make the following matrix transformation,

$$PAP^{-1} = \begin{bmatrix} \bar{A}_1 & 0 \\ \bar{A}_2 & A_{44} \end{bmatrix}, CP^{-1} = \left[\begin{array}{cc} \bar{C} & 0 \end{array}\right],$$

where $\bar{A}_1 = M_{n,\Sigma}A\left[M_{n,\Sigma}^+ \quad \bar{V}\right]$, $\bar{A}_2 = V^+A\left[M_{n,\Sigma}^+ \quad \bar{V}\right]$, and $\bar{C} = C\left[M_{n,\Sigma}^+ \quad \bar{V}\right]$. It is known that, for this kind of transformation, the pair $(A_1,\bar{C})$ is observable (see, e.g., [26]), and the (A, C)-unobservable eigenvalues are the eigenvalues of the matrix A_{44}, which proves the clause a). Besides, as it was established in Lemma 3.7, the invariant zeros of (A, C, D, F) are the eigenvalues of A_4. Therefore, taking into account the specific form of A_4 obtained in (4.4), the set of invariant zeros of Σ that do not belong to the set of (A, C)-unobservable eigenvalues is identical to the set of eigenvalues of A_{41}, which proves Lemma b). □

PROPOSITION A.1. *Under the condition $\mathcal{N}^* \neq \mathcal{V}_\Sigma^*$, for any matrices V and K_1^* satisfying* (4.2) *and* (4.3), *respectively, the pair (A_{41}, K_1^*) is observable.*

Proof of Proposition A.1. Suppose that (A_{41},K_1^*) is unobservable, then there is a vector $p \neq 0$ so that $A_{41}p = \lambda p$ and $K_1^*p = 0$ for some scalar constant λ. Then, since $A_{41} = Q_1$, from (3.6), (4.2), and (4.3), we have

$$\begin{aligned} A\left[\begin{array}{cc} \bar{V}p & N \end{array}\right] &= \left[\begin{array}{cc} \bar{V}p & N \end{array}\right]\begin{bmatrix} \lambda & 0 \\ Q_2p & Q_4 \end{bmatrix} \\ C\left[\begin{array}{cc} \bar{V}p & N \end{array}\right] &= 0. \end{aligned}$$

That is, the span of $\left[\begin{array}{cc} \bar{V}p & N \end{array}\right]$ is an A-invariant subspace with dimension bigger than $\mathcal{N}^*$, which is a contradiction since $\mathcal{N}^*$ is the greatest A-invariant subspace belonging to $\ker C$. □

Proof of Theorem 4.7. Firstly, let us prove the equivalence (iii)⇔(iv). If (iii) is true, from Lemma 4.2 $\operatorname{rank}\begin{bmatrix} D_1^T & F^T \end{bmatrix}^T = q$. Furthermore, since $\ker M_{n,\Sigma} = \mathcal{V}^* = \mathcal{N}^* = \ker O$, we have $\operatorname{rank} M_{n,\Sigma} = \operatorname{rank} O$. On the other hand, if (iv) is true, then

$$\dim \mathcal{V}^* = \dim \ker M_{n,\Sigma} = n - \operatorname{rank} M_{n,\Sigma} = n - \operatorname{rank} O = \dim \ker O = \dim \mathcal{N}^*$$

Therefore, $\mathcal{V}^* = \mathcal{N}^*$. The previous identity and Lemma 4.2 prove the implication ⇐).

The proof of (ii)⇔(iv) is as follows. Supposing that (iv) is true, since $V = N$, $P^{-1} = \begin{bmatrix} M_{n,\Sigma}^{+} & N \end{bmatrix}$ and $A_4 = A_{44}$. Therefore, the set of eigenvalues of A_4 is at the same time the set of the invariant zeros of Σ and the set of (A, C)-unobservable eigenvalues. Thus, ⇐) is proven. Now, suppose that (ii) is satisfied. Then, by Lemma 3.7.b), $\operatorname{rank}\begin{bmatrix} D_1^T & F^T \end{bmatrix}^T = q$. Moreover, by Lemma 3.7.a) and Lemma 4.6.a), $A_4 = A_{44}$, i.e., $\dim \mathcal{V}^* = n - n_1 - n_2 = \dim \mathcal{N}^*$, which implies $\operatorname{rank} M_{n,\Sigma} = \operatorname{rank} O$. Thus, the implication ⇒) is proven.

Now, let us prove the equivalence (i)⇔(iv). Suppose $\mathcal{V}^* \neq \mathcal{N}^*$, then by choosing $x_1(0) = 0$ and $w(t) = k_1^* \bar{x}_{21}(t)$, we obtain the identities $x_1(t) \equiv 0$, $y(t) \equiv 0$, and $\dot{\bar{x}}_{21}(t) = A_{41}\bar{x}_{21}(t)$. Thus, for the proposition (A.1), if $\bar{x}_{21}(0) \neq 0$, $w \neq 0$. This means that for $\mathcal{V}^* \neq \mathcal{N}^*$ there exist the conditions such that $w(t) \neq 0$ in spite of $y(t) \equiv 0$. Therefore, the claim (i) is achieved only if the identity $\mathcal{V}^* = \mathcal{N}^*$ is true, that is, if $\operatorname{rank} M_{n,\Sigma} = \operatorname{rank} O$. Furthermore, it is clear that the left invertibility property is necessary for fulfilling (4.6). Therefore, by Lemma 4.2 we have that $\operatorname{rank}\begin{bmatrix} D_1^T & F^T \end{bmatrix}^T = q$, and (i)⇒(iv) is proven. Now, suppose that (iv) is true and $y(t) \equiv 0$. Then, we have, from Lemma 3.4, $\bar{x}_1(t) \equiv 0$ and $\bar{w}(t) \equiv 0$. But, because of in this case $V = N$, then $w(t) = \bar{w}(t) \equiv 0$. Therefore, (iv)⇒(i) is proven. □

Part of the proof of Theorem 4.8 is based on the following proposition.

Proposition A.2. *The set of the invariant zeros of* Σ *that do not belong to the set of* (A, C)*-unobservable eigenvalues is in* $\mathbb{C}^-$ *if, and only if,*

$$\operatorname{rank}\begin{bmatrix} D_1 \\ F \end{bmatrix} = q \text{ and } A_{41} \text{ is Hurwitz.} \tag{A.4}$$

Proof of Proposition A.2. Suppose $\operatorname{rank}\left[\begin{smallmatrix} D_1 \\ F \end{smallmatrix}\right] < q$, then any $s \in \mathbb{C}$ is an invariant zero of Σ (Lemma 3.7). Hence, since the set of (A, C)-unobservable eigenvalues is finite, in this case, there is a set (infinite) of invariant zeros of Σ that do not belong to the set of (A, C)-unobservable eigenvalues having positive real part. Thus, we have proven that $\operatorname{rank}\left[\begin{smallmatrix} D_1 \\ F \end{smallmatrix}\right] = q$, which implies, due to Lemma 4.6, that A_{41} is a Hurwitz matrix.

The sufficiency comes from (A.4) and Proposition 4.6.b). □

Proof of Theorem 4.8. The equivalence (ii)⇔(iii) follows directly from Proposition A.2.

Now, suppose that the clause (i) is true. From the proof of necessity of Lemma 4.2, we have that the condition $\operatorname{rank}\left[\begin{smallmatrix} D_1 \\ F \end{smallmatrix}\right] = q$ is essential for fulfilling the clause (i). Now, selecting $\bar{x}_1(0) = 0$ and $w(t) = K_{21}^* \bar{x}_{21}$, we obtain the identities $\bar{x}_1(t) \equiv 0$, $\dot{\bar{x}}_{21}(t) = A_{41}\bar{x}_{21}(t)$, and $y(t) \equiv 0$; therefore, $w(t) \to 0$. Now, if $\mathcal{V}^* = \mathcal{N}^*$, the set of eigenvalues of A_{41} is empty. If $\mathcal{V}^* \neq \mathcal{N}^*$, by Proposition A.1, $w(t)$ tends to zero if, and only if, $\bar{x}_{21}(t)$ tends to zero. Hence, we conclude that A_{41} is Hurwitz. Thus, we have the implication (i)⇒(iii).

Now, suppose that (iii) is true. If $y(t) \equiv 0$, we have $\bar{x}_1(t) \equiv 0$, $\dot{\bar{x}}_{21}(t) = A_{41}\bar{x}_{21}(t)$, and $w(t) = K_1^* \bar{x}_{21}(t)$. Since A_{41} is Hurwitz, it means $\bar{x}_{21}(t) \to 0$ and so $w(t) \to 0$. Thus, the implication (i)⇐(iii) is proven. □

Proof of Theorem 4.9. Evidently, the implication (4.7) is a necessary condition for the estimate of w. Hence, the necessity follows from Theorem 4.8. Now, suppose that the invariant zeros of Σ not belonging to the set of (A, C)-unobservable eigenvalues have negative real part. Then, from the proposition A.2, rank $\left[\begin{smallmatrix} D_1 \\ F \end{smallmatrix}\right] = q$ and all the eigenvalues of A_{41} have negative real part. Thus, from (4.5), $\bar{w}(t)$ can be expressed by the following equation:

$$\bar{w}(t) = \begin{bmatrix} D_1 \\ F \end{bmatrix}^+ \begin{bmatrix} \dot{\bar{x}}_1(t) - A_1 \bar{x}_1(t) \\ y - C_1 \bar{x}_1(t) \end{bmatrix} \tag{A.5}$$

Since $\bar{x}_1$ always can be reconstructed using (3.11), also $\bar{w}(t)$ can be reconstructed. Thus, estimating $w = \bar{w}(t) + K_1^* \bar{x}_{21}$ is equivalent to estimating $K_1^* \bar{x}_{21}$. Hence, substituting $\bar{w}$, given by (A.5), into the dynamic equation of $\bar{x}_{21}$, and defining

$$\begin{aligned} \dot{\tilde{z}}_{21} &= A_{41}\hat{z}_{21} + A_{21}\bar{x}_1 - D_{21}\begin{bmatrix} D_1 \\ F \end{bmatrix}^+ \begin{bmatrix} A_1\bar{x}_1 \\ C_1\bar{x}_1 - y \end{bmatrix} \\ \hat{z}_{21} &= \tilde{z}_{21} + D_{21}\begin{bmatrix} D_1 \\ F \end{bmatrix}^+ \begin{bmatrix} \bar{x}_1 \\ 0 \end{bmatrix}, \end{aligned} \tag{A.6}$$

we obtain that dynamic equation for the error $e_{21} := \bar{x}_{21} - \hat{z}_{21}$ is $\dot{e}_{21}(t) = A_{41}\dot{e}_{21}(t)$. Thus, since the eigenvalues of A_{41} have negative real part, then $\hat{z}_{21}(t) \to \bar{x}_{21}(t)$ as $t \to \infty$. Obviously $K_1^* \hat{z}_{21}(t) \to K_1^* \bar{x}_{21}(t)$ (as $t \to \infty$). This completes the proof. □

Proof of Theorem 4.10. The necessity comes from (4.6) and Theorem 4.7. On the other hand, if the set of the invariant zeros of Σ is identical to the set of (A, C)-unobservable eigenvalues, then $V = N$ and the identity $w(t) = \bar{w}(t)$ holds. Therefore, w can be reconstructed directly from (A.5) with $\bar{x}_1$ obtained from (3.11). □

Acknowledgments. The authors thank Dr. Christopher Edwards, Dr. Thierry Floquet, and Prof. Mehrdad Saif for their fruitful discussions. The work of L. Fridman was supported by the Mexican CONACYT under the Grant 56819, the PAPIIT (Programa de Apoyo a Proyectos de Investigación e Inovación Tecnológica) UNAM under Grant 111208, and the PAPIME (Programa de Apoyo a Proyectos Institucionales para el Mejoramiento de la Enseñanza) PE 100907. The work of F.J. Bejarano was supported by the Mexican CONACYT postdoctoral grant, CVU 103957.

REFERENCES

[1] T. Ahmed-Ali and F. Lamnabhi-Lagarrigue, *Sliding observer-controller design for uncertain triangular nonlinear systems*, IEEE Trans. Automat. Control, 44 (1999), pp. 1244–1249.

[2] J. Alvarez, Y. Orlov, and L. Acho, *An invariance principle for discontinuous dynamic systems with application to a coulomb friction oscillator*, J. Dynamic Systems, Measurement, and Control, 122 (2000), pp. 687–690.

[3] J. Barbot, M. Djemai, and T. Boukhobza, *Sliding mode observers*, in Sliding Mode Control in Engineering, W. Perruquetti and J. Barbot, eds., Control Engineering, Marcel Dekker, New York, 2002, pp. 103–130.

[4] G. Bartolini, A. Levant, A. Pisano, and E. Usai, *Higer-order sliding modes for the output-feedback control of nonlinear uncertain systems*, in Variable Structure Systems: Towards the 21st Century, X.Yu and J.-X.Xu, eds., Lecture Notes in Control and Information Science, Springer Verlag, Berlin, 2002, pp. 83–108.

[5] G. Bartolini, A. Pisano, E. Punta, and E. Usai, *A survey of applications of second-order sliding mode control to mechanical systems*, Int. J. Control, 76 (2003), pp. 875–892.

[6] G. Basile and G. Marro, *Controlled and Conditioned Invariants in Linear System Theory*, Prentice Hall, Englewood Cliffs, NJ, 1992.

[7] F. Bejarano, L. Fridman, and A. Poznyak, *Exact state estimation for linear systems with unknown inputs based on hierarchical super-twisting algorithm*, Int. J. Robust and Nonlinear Control, 17 (2007), pp. 1734–1753.

[8] F. Bejarano, A. Poznyak, and L. Fridman, *Hierarchical second-order sliding mode observer for linear time invariant systems with unknown inputs*, Int. J. Systems Sciences, 38 (2007), pp. 793–802.

[9] J. Davila, L. Fridman, and A. Poznyak, *Observation and identification of mechanical systems via second order sliding modes*, Int. J. Control, 79 (2006), pp. 1251–1262.

[10] C. Edwards, S. Spurgeon, and R. Hebden, *On development and applications of sliding mode observers*, in Variable Structure Systems: Towards XXIst Century, J. Xu and Y. Xu, eds., Lecture Notes in Control and Information Science, Springer Verlag, Berlin, Germany, 2002, pp. 253–282.

[11] C. Edwards and S. Spurgeon, *Sliding Mode Control*, Taylor and Francis, London, 1998.

[12] T. Floquet and J. Barbot, *A canonical form for the design of unknown imput sliding mode observers*, in Advances in Variable Structure and Sliding Mode Control, C. Edwards, E. Fossas, and L. Fridman, eds., Lecture Notes in Control and Information Sciences 334, Springer Verlag, Berlin, 2006, pp. 271–292.

[13] T. Floquet, C. Edwards, and S. Spurgeon, *On sliding mode observers for systems with unknown inputs*, Int. J. Adapt. Control Signal Process, 21 (2007), pp. 638–656.

[14] L. Freidovich and H. Khalil, *Lyapunov-based switching control of nonlinear systems using high-gain observers*, Automatica, 43 (2007), pp. 150–157.

[15] H. Hashimoto, V. Utkin, J. Xu, H. Suzuki, and F. Harashima, *Vss observer for linear time varying system*, in Procedings of IECON'90, Pacific Grove, CA, 1990, pp. 34–39.

[16] M. Hautus, *Strong detectability and observers*, Linear Algebra Appl., 50 (1983), pp. 353–368.

[17] A. Levant, *Sliding order and sliding accuracy in sliding mode control*, Int. J. Control, 58 (1993), pp. 1247–1263.

[18] A. Levant, *Robust exact differentiation via sliding mode technique*, Automat., 34 (1998), pp. 379–384.

[19] A. Levant, *High-order sliding modes: Differentiation and output-feedback control*, Int. J. Control, 76 (2003), pp. 924–941.

[20] B. Molinari, *A strong contollability and observability in linear multivariable control*, IEEE Trans. Automat. Control, 21 (1976), pp. 761–764.

[21] Y. Orlov, L. Aguilar, and J. Cadiou, *Switched chattering control vs. backlash/friction phenomena in electrical servo-motors*, Int. J. Control, 76 (2003), pp. 959–967.

[22] A. Poznyak, *Deterministic output noise effects in sliding mode observation*, in Variable structure systems: from principles to implementation, A. Sabanovic, L. Fridman, and S. Spurgeon, eds., IEEE Control Engineering Series, IEEE, London, 2004, pp. 45–80.

[23] M. Saif and Y. Xiong, *Sliding mode observers and their application in fault diagnosis*, in Fault Diagnosis and Fault Tolerance for Mechatronic Systems: Recent Advances, F. Caccavale and L. Villani, eds., vol. 1/2003 of Springer Tracts in Advanced Robotics, Springer, Berlin, 2003, pp. 1–57.

[24] Y. Shtessel, I. Shkolnikov, and M. Brown, *An asymptotic second-order smooth sliding mode control*, Asian J. Control, 5 (2003), pp. 498–504.

[25] C. Tan and C. Edwards, *Sliding mode observers for robust detection and reconstruction of actuator and sensor faults*, Int. J. Robust Nonlinear Control, 13 (2003), pp. 443–463.

[26] H. Trentelman, A. Stoorvogel, and M. Hautus, *Control Theory for Linear Systems*, Communications and control engineering, Springer, New York, London, 2001.

[27] V. Utkin, J. Guldner, and J. Shi, *Sliding Modes in Electromechanical Systems*, Taylor and Francis, London, 1999.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 2, pp. 1179–1205

© 2009 Society for Industrial and Applied Mathematics

PRACTICAL STABILIZATION OF A QUANTUM PARTICLE IN A ONE-DIMENSIONAL INFINITE SQUARE POTENTIAL WELL*

KARINE BEAUCHARD† AND MAZYAR MIRRAHIMI‡

Abstract. We consider a nonrelativistic charged particle in a one-dimensional infinite square potential well. This quantum system is subjected to a control, which is a uniform (in space) time-depending electric field. It is represented by a complex probability amplitude solution of a Schrödinger equation on a one-dimensional bounded domain, with Dirichlet boundary conditions. We prove the almost global practical stabilization of the eigenstates by explicit feedback laws.

Key words. control of partial differential equations, bilinear Schrödinger equation, quantum systems, Lyapunov stabilization

AMS subject classifications. 93C20, 35Q40, 93D15

DOI. 10.1137/070704204

1. Introduction.

1.1. Main result.

As in [23, 5, 6], we consider a nonrelativist charged particle in a one-dimensional space, with a potential $V(x)$, in a uniform electric field $t \mapsto u(t) \in \mathbb{R}$. Under the dipole moment approximation assumption, and after appropriate changes of scales, the evolution of the particle's wave function is given by the following Schrödinger equation:

$$i\frac{\partial \Psi}{\partial t}(t,x) = -\frac{1}{2}\frac{\partial^2 \Psi}{\partial x^2}(t,x) + (V(x) - u(t)x)\Psi(t,x).$$

We study the case of an infinite square potential well: $V(x) = 0$ for $x \in I := (-1/2, 1/2)$ and $V(x) = +\infty$ for x outside I. Therefore our system is

$$i\frac{\partial \Psi}{\partial t}(t,x) = -\frac{1}{2}\frac{\partial^2 \Psi}{\partial x^2}(t,x) - u(t)x\Psi(t,x), \qquad x \in I, \tag{1.1}$$

$$\Psi(0,x) = \Psi_0(x), \tag{1.2}$$

$$\Psi(t, \pm 1/2) = 0. \tag{1.3}$$

It is a nonlinear control system, denoted by (Σ), in which

- the control is the electric field $u : \mathbb{R}_+ \to \mathbb{R}$;
- the state is the wave function $\Psi : \mathbb{R}_+ \times I \to \mathbb{C}$ with $\Psi(t) \in \mathbb{S}$ for every $t \geqslant 0$, where $\mathbb{S} := \{\varphi \in L^2(I;\mathbb{C}); \|\varphi\|_{L^2} = 1\}$.

Let us introduce the operator A defined by

$$D(A) := (H^2 \cap H^1_0)(I, \mathbb{C}), \quad A\varphi := -\frac{1}{2}\frac{d^2\varphi}{dx^2},$$

*Received by the editors October 1, 2007; accepted for publication (in revised form) December 14, 2008; published electronically March 18, 2009. This work was supported in part by the "Agence Nationale de la Recherche" (ANR), Projet Blanc CQUID 06-3-13957. A short and preliminary version of this paper was submitted to the IFAC World Congress 2008.
http://www.siam.org/journals/sicon/48-2/70420.html

†CNRS, CMLA, ENS Cachan, Avenue du président Wilson, 94230 Cachan, France (karine.beauchard@cmla.ens-cachan.fr).

‡INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt B.P. 105, 78153 Le Chesnay Cedex, France (mazyar.mirrahimi@inria.fr).

and for $s \in \mathbb{R}$ the spaces

$$H^s_{(0)}(I,\mathbb{C}) := D(A^{s/2}).$$

The following proposition recalls classical existence and uniqueness results for the solutions of (1.1)–(1.3). For sake of completeness, a proof of this proposition is given in the appendix.

Proposition 1.1. *Let $\Psi_0 \in \mathbb{S}$, $T > 0$, and $u \in C^0([0,T],\mathbb{R})$. There exists a unique weak solution of* (1.1)–(1.3), *i.e., a function $\Psi \in C^0([0,T],\mathbb{S}) \cap C^1([0,T], H^{-2}_{(0)}(I,\mathbb{C}))$ such that*

$$\Psi(t) = e^{-iAt}\Psi_0 + i\int_0^t e^{-iA(t-s)}u(s)x\Psi(s)ds \textit{ in } L^2(I,\mathbb{C}) \textit{ for every } t \in [0,T]. \tag{1.4}$$

Then (1.1) *holds in $H^{-2}_{(0)}(I,\mathbb{C})$ for every $t \in [0,T]$.*

If, moreover, $\Psi_0 \in (H^2 \cap H^1_0)(I,\mathbb{C})$, then Ψ is a strong solution, i.e., $\Psi \in C^0([0,T],(H^2 \cap H^1_0)(I,\mathbb{C})) \cap C^1([0,T],L^2(I,\mathbb{C}))$, (1.1) *holds in $L^2(I,\mathbb{C})$ for every $t \in [0,T]$,* (1.2) *holds in $H^2 \cap H^1_0(I,\mathbb{C})$, and* (1.3) *holds for every $t \in [0,T]$.*

The weak (resp., strong) solutions are continuous with respect to initial conditions for the $C^0([0,T],L^2)$-topology (resp., for the $C^0([0,T],H^2 \cap H^1_0)$-topology.)

The symbol $\langle .,. \rangle$ denotes the usual Hermitian product of $L^2(I,\mathbb{C})$, i.e.,

$$\langle \varphi, \xi \rangle := \int_I \varphi(x)\overline{\xi(x)}dx.$$

For $\sigma \in \mathbb{R}$, we introduce the operator A_σ defined by

$$D(A_\sigma) := (H^2 \cap H^1_0)(I;\mathbb{C}), \quad A_\sigma\varphi := -\frac{1}{2}\frac{\partial^2\varphi}{dx^2} - \sigma x\varphi.$$

It is well known that there exists an orthonormal basis $(\phi_{k,\sigma})_{k\in\mathbb{N}^*}$ of $L^2(I,\mathbb{C})$ of eigenvectors of A_σ:

$$\phi_{k,\sigma} \in H^2 \cap H^1_0(I,\mathbb{C}), \quad A_\sigma\phi_{k,\sigma} = \lambda_{k,\sigma}\phi_{k,\sigma},$$

where $(\lambda_{k,\sigma})_{k\in\mathbb{N}^*}$ is a nondecreasing sequence of real numbers. For $s > 0$ and $\sigma \in \mathbb{R}$, we define

$$H^s_{(\sigma)}(I,\mathbb{C}) := D(A_\sigma^{s/2}),$$

equipped with the norm

$$\|\varphi\|_{H^s_{(\sigma)}} := \left(\sum_{k=1}^{\infty} \lambda^s_{k,\sigma}|\langle \varphi, \phi_{k,\sigma}\rangle|^2\right)^{1/2}.$$

For $k \in \mathbb{N}^*$ and $\sigma \in \mathbb{R}$, we define

$$\mathcal{C}_{k,\sigma} := \{\phi_{k,\sigma}e^{i\theta}; \theta \in [0,2\pi)\}.$$

In order to simplify the notation, we will write ϕ_k, λ_k, $\mathcal{C}_k$ instead of $\phi_{k,0}$, $\lambda_{k,0}$, $\mathcal{C}_{k,0}$. We have

$$\lambda_k = \frac{k^2\pi^2}{2}, \quad \phi_k = \begin{cases} \sqrt{2}\cos(k\pi x) \text{ when } k \text{ is odd,} \\ \sqrt{2}\sin(k\pi x) \text{ when } k \text{ is even.} \end{cases} \tag{1.5}$$

The goal of this paper is the study of the stabilization of the system (Σ) around the eigenstates $\phi_{k,\sigma}$. More precisely, for $k \in \mathbb{N}^*$ and $\sigma \in \mathbb{R}$ small, we state feedback laws $u = u_{k,\sigma}(\Psi)$ for which the solution of (1.1)–(1.3) with $u(t) = u_{k,\sigma}(\Psi(t))$ is such that

$$\limsup_{t\to+\infty} \mathrm{dist}_{L^2(I,\mathbb{C})}(\Psi(t), \mathcal{C}_{k,\sigma})$$

is arbitrarily small. We consider convergence toward the circle $\mathcal{C}_{k,\sigma}$ because the wave function Ψ is defined up to a phase factor. For simplicity's sake, we will only work with the ground state $\phi_{1,\sigma}$. However, the whole argument remain valid for the general case.

Note that even though the feedback stabilization of a quantum system necessitates more complicated models taking into account the measurement back action on the system (see, e.g., [14, 29, 19]), the kind of strategy considered in this paper can be helpful for the open-loop control of closed quantum systems. Indeed, one can apply the stabilization techniques for the Schrödinger equation in simulation and retrieve the control signal that will then be applied in open-loop on the real physical system. As will be shown in detail below, in the bibliographic overview, this kind of strategy has been widely used in the context of finite-dimensional quantum systems.

The main result of this article is the following.

THEOREM 1.2. *Let* $\Gamma > 0$, $s > 0$, $\epsilon > 0$, $\gamma \in (0, 1)$. *There exists* $\sigma^{**} = \sigma^{**}(\Gamma, s) > 0$ *such that, for every* $\sigma \in (-\sigma^{**}, \sigma^{**})$, *there exists a feedback law* $v_{\sigma,\Gamma,s,\epsilon,\gamma}(\Psi)$ *such that, for every* $\Psi_0 \in \mathbb{S} \cap (H^2 \cap H_0^1 \cap H^s_{(\sigma)})(I, \mathbb{C})$ *with*

$$\|\Psi_0\|_{H^s_{(\sigma)}} \leqslant \Gamma \text{ and } |\langle \Psi_0, \phi_{1,\sigma}\rangle| > \gamma,$$

the Cauchy problem (1.1)–(1.3) *with* $u(t) = \sigma + v_{\sigma,\Gamma,s,\epsilon,\gamma}(\Psi)$ *has a unique strong solution; moreover, this solution satisfies*

$$\limsup_{t\to+\infty} dist_{L^2}(\Psi(t), \mathcal{C}_{1,\sigma}) \leqslant \epsilon.$$

Remark 1. Theorem 1.2 provides *almost global practical stabilization.* In fact, as will be seen, the above feedback law may be found through a Lyapunov analysis which ensures the stability of the target; i.e., for any small $\epsilon_1 > 0$, there exists a $0 < \epsilon_2 < \epsilon_1$ such that, if we initialize the system in an ϵ_2-neighborhood, the solution does not get outside the ϵ_1-neighborhood.

Moreover, applying the theorem, any initial condition $\Psi_0 \in \mathbb{S}$ such that $\Psi_0 \in H^s(I, \mathbb{C})$ for some $s > 0$ and $\langle \Psi_0, \phi_{1,\sigma}\rangle \neq 0$ can be moved approximately to the circle $\mathcal{C}_{1,\sigma}$, thanks to an appropriate feedback law.

The stability and the approximate convergence lead to the practical stabilization.

For $\sigma \neq 0$, the feedback law will be given explicitly. For $\sigma = 0$, the feedback law will be given by an implicit formula. We will see that the assumption "$\Psi_0 \in H^s(I, \mathbb{C})$ for some $s > 0$" is not necessary for the result of the theorem. In fact, even for a Ψ_0 only belonging to $\mathbb{S}$, we can find the appropriate feedback law as a function of the initial state Ψ_0.

Notice that, physically, the assumption $\langle \Psi_0, \phi_{1,\sigma}\rangle \neq 0$ is not really restrictive. Indeed, if $\langle \Psi_0, \phi_{1,\sigma}\rangle = 0$, a control field in resonance with the natural frequencies of the system (the difference between the eigenvalues corresponding to an eigenstate whose population in the initial state is nonzero and the ground state) will, instantaneously, ensure a nonzero population of the ground state in the wavefunction. Then one can just apply the feedback law of Theorem 1.2.

1.2. A brief bibliography. The controllability of a finite-dimensional quantum system, $\iota \frac{d}{dt}\Psi = (H_0 + u(t)\ H_1)\Psi$, where $\Psi \in \mathbb{C}^N$ and H_0 and H_1 are $N \times N$ Hermitian matrices with coefficients in $\mathbb{C}$, has been very well explored [26, 22, 1, 2, 28]. However, this does not guarantee the simplicity of the trajectory generation. Very often the chemists formulate the task of the open-loop control as a cost functional to be minimized. Optimal control techniques (see, e.g., [24]) and iterative stochastic techniques (e.g., genetic algorithms [17]) are then two classes of approaches which are most commonly used for this task.

When some nondegeneracy assumptions concerning the linearized system are satisfied, [20] provides another method based on Lyapunov techniques for generating trajectories. The relevance of such a method for the control of chemical models has been studied in [21]. As mentioned above, the closed-loop system is simulated and the retrieved control signal is applied in open-loop. Such a strategy has already been applied widely in this framework [8, 25].

The situation is much more difficult when we consider an infinite-dimensional configuration, and fewer results are available. However, the controllability of the system (1.1)–(1.3) is now well understood. In [27], the author states some noncontrollability results for general Schrödinger systems. These results apply in particular to the system (1.1)–(1.3). However, this negative result is due to the choice of the functional space that does not allow controllability. Indeed, if we consider different functional spaces, one can get positive controllability results. In [5], the local controllability of the system (1.1)–(1.3) around the ground state $\phi_{1,\sigma}$ for σ small is proved. The case $\sigma \neq 0$ is easier because the linearized system around $\phi_{1,\sigma}$ for $\sigma \neq 0$ small is controllable; this case is treated with the moment theory and a Nash–Moser implicit functions theorem. As has been discussed in [23], the case of $\sigma = 0$ is degenerate: the linearized system around ϕ_1 is not controllable. Therefore, in this case, one needs to apply other tools, namely, the return method (introduced in [9]) and the quantum adiabatic theory [12]. In [6], the steady-state controllability of this nonlinear system is proved (i.e., the particle can be moved in finite time from an eigenstate ϕ_k to another one, ϕ_j). The proof relies on many local controllability results (proved with the previous strategies) together with a compactness argument.

Concerning the trajectory generation problem for infinite-dimensional systems, still much fewer results are available. What literature exists is mostly based on the use of the optimal control techniques [4, 3]. The simplicity of the feedback law found by the Lyapunov techniques in [20, 7] suggests the use of the same approach for infinite-dimensional configurations. However, an extension of the convergence analysis to the PDE configuration is not at all a trivial problem. Indeed, it requires the precompactness of the closed-loop trajectories, a property that is difficult to prove in infinite dimension. This strategy is used, for example, in [11].

In [18], one of the authors proposes a Lyapunov-based method for practical stabilization of a particle in an N-dimensional decaying potential under some restrictive assumptions. The author assumes that the system is initialized in the finite-dimensional discrete part of the spectrum. Then the idea consists in proposing a Lyapunov function which encodes both the distance with respect to the target state and the necessity of remaining in the discrete part of the spectrum. In this way, he prevents the possibility of the "mass lost phenomenon" through dispersion at infinity. Applying some dispersive estimates of Strichartz type, he ensures the practical stabilization of an arbitrary eigenstate in the discrete part of the spectrum.

Finally, let us mention that there exists a huge literature on the other strategies for proving the stabilization of infinite-dimensional control systems. We refer to [10] for a rather complete list of references on these techniques.

In this paper, we study the stabilization of the ground state $\phi_{1,\sigma}$ for σ in a neighborhood of 0. Adapting the techniques proposed in [18], we ensure the practical stabilization of the system around $\phi_{1,\sigma}$. Note that the whole argument holds if we replace the target state by any eigenstate $\phi_{k,\sigma}$ of the system.

1.3. Heuristic of the proof. To stabilize the ground state $\phi_{1,\sigma}$, a first approach would be to consider the simple Lyapunov function

$$\widetilde{\mathcal{V}}(\Psi) = 1 - |\langle \Psi \ , \ \phi_{1,\sigma} \rangle|^2.$$

Just as in the finite-dimensional case [7], the feedback law

$$\tilde{u}(\Psi) = \Im(\langle x\Psi \ , \ \phi_{1,\sigma} \rangle \langle \phi_{1,\sigma} \ , \ \Psi \rangle),$$

where $\Im$ denotes the imaginary part of a complex, ensures the decrease of the Lyapunov function. However, trying to adapt the convergence analysis, based on the use of the LaSalle invariance principle, the precompactness of the trajectories in L^2 constitutes a major obstacle. Note that in order to be able to apply the LaSalle principle to an infinite-dimensional system, one certainly needs to prove such a precompactness result. For the particular case of the infinite potential well considered here, the efforts of the authors, applying the classical functional analysis techniques, have failed to prove the precompactness of the closed-loop system applying the above feedback. In fact, as the system evolves on the unit sphere of L^2, the compactness of the trajectories in weaker spaces is ensured. However, we have not been able to strengthen this weak compactness to a strong one. Indeed, it even seems that phenomena such as the loss of L^2-mass in the high energy levels do not allow this property to hold true.

Similarly to [18], the approach of this paper is to prevent the population from going through the very high energy levels, while trying to stabilize the system around $\phi_{1,\sigma}$.

As in Theorem 1.2, let us consider $\Gamma > 0$, $s > 0$, $\epsilon > 0$, $\gamma > 0$, $\sigma \in \mathbb{R}$. First, we consider the case $\sigma \neq 0$. Let $\Psi_0 \in H^s_{(0)}(I, \mathbb{C})$ with

$$\|\Psi_0\|_{H^s_{(0)}} \leqslant \Gamma \text{ and } |\langle \Psi_0, \phi_{1,\sigma} \rangle| \geqslant \gamma.$$

We claim that there exists $N = N(\Gamma, s, \epsilon, \gamma) \in \mathbb{N}^*$ large enough so that

$$\sum_{k=N+1}^{\infty} |\langle \Psi_0, \phi_{k,\sigma} \rangle|^2 < \frac{\epsilon\gamma^2}{1-\epsilon}. \tag{1.6}$$

Then we consider the Lyapunov function

$$\mathcal{V}(\Psi) = 1 - |\langle \Psi \ , \ \phi_{1,\sigma} \rangle|^2 - (1-\epsilon) \sum_{k=2}^{N} |\langle \Psi \ , \ \phi_{k,\sigma} \rangle|^2. \tag{1.7}$$

Note that this Lyapunov function depends on the constants Γ, s, ϵ, γ through the choice of the cut-off dimension, N. Just like [18], it encodes two tasks: (1) it prevents the loss of L^2-mass through the high-energy eigenstates; and (2) it privileges the increase of the population in the first eigenstate.

When Ψ solves (Σ) with some control $u = \sigma + v$, we have

$$\frac{d\mathcal{V}}{dt} = -2v(t)\Im\Big(\sum_{k=1}^{N} a_k \left\langle x\Psi \ , \ \phi_{k,\sigma}\right\rangle \left\langle \phi_{k,\sigma} \ , \ \Psi\right\rangle\Big),$$

where

$$a_1 := 1 \quad \text{and} \quad a_k := 1-\epsilon \text{ for } k = 2, \dots, N. \tag{1.8}$$

Thus, the feedback law

$$v(\Psi) := \varsigma\Im\Big(\sum_{k=1}^{N} a_k \left\langle x\Psi \ , \ \phi_{k,\sigma}\right\rangle \left\langle \phi_{k,\sigma} \ , \ \Psi\right\rangle\Big), \tag{1.9}$$

where $\varsigma > 0$ is a positive constant, trivially ensures the decrease of the Lyapunov function (1.7). We claim that the solution of (1.1)–(1.3) with initial condition Ψ_0 and control $u = \sigma + v(\Psi)$ satisfies

$$\limsup_{t\to+\infty} \operatorname{dist}_{L^2}(\Psi(t), \mathcal{C}_{1,\sigma})^2 \leqslant \epsilon. \tag{1.10}$$

Note that the claimed result here is much stronger than the one provided in [18] for the finite potential well problem. In fact, here we claim the almost global practical stabilization of the system round the eigenstate $\phi_{1,\sigma}$.

The limit (1.10) will be proved by studying the $L^2(I,\mathbb{C})$-weak limits of $\Psi(t)$ when $t \to +\infty$. Namely, let $(t_n)_{n\in\mathbb{N}}$ be an increasing sequence of positive real numbers such that $t_n \to +\infty$. Since $\|\Psi(t_n)\|_{L^2(I,\mathbb{C})} \equiv 1$, there exists $\Psi_\infty \in L^2(I,\mathbb{C})$ such that, up to a subsequence, $\Psi(t_n) \to \Psi_\infty$ weakly in $L^2(I,\mathbb{C})$. Using the controllability of the linearized system around $\phi_{1,\sigma}$ (which is equivalent to $\langle \phi_{1,\sigma}, x\phi_{k,\sigma}\rangle \neq 0$ for every $k \in \mathbb{N}^*$), we will be able to prove that $\Psi_\infty = \beta\phi_{1,\sigma}$, where $\beta \in \mathbb{C}$ and $|\beta|^2 \geqslant 1-\epsilon$. This will imply (1.10).

Therefore, by weakening the stabilization property (i.e., looking for practical stabilization instead of stabilization) we avoid the compactness problem evoked at the beginning of this section.

Note that the controllability of the linearized system around the trajectory $\phi_{1,\sigma}$ plays a crucial role here. This is why the developed techniques for $\sigma \neq 0$ cannot be applied, directly, to the case of $\sigma = 0$.

Now, let us study the case $\sigma = 0$. As emphasized above, the previous strategy does not work for the practical stabilization of ϕ_1 because the linearized system around ϕ_1 is not controllable. The idea is thus to use the above feedback design (1.9) with a dynamic $\sigma = \sigma(t)$ that converges to zero as $t \to +\infty$. Formally, the convergence of Ψ toward $\mathcal{C}_{1,\sigma(t)}$ must happen at a faster rate than that of σ toward zero (see Figure 1.1).

In this aim, we consider the Lyapunov function

$$\mathcal{V}(\Psi) = 1 - (1-\epsilon)\sum_{k=1}^{N} \left| \left\langle \Psi \ , \ \phi_{k,\sigma(\Psi)}\right\rangle \right|^2 - \epsilon \left| \left\langle \Psi \ , \ \phi_{1,\sigma(\Psi)}\right\rangle \right|^2, \tag{1.11}$$

where the function $\Psi \mapsto \sigma(\Psi)$ is implicitly defined as

$$\sigma(\Psi) = \theta\left(\mathcal{V}(\Psi)\right) \tag{1.12}$$

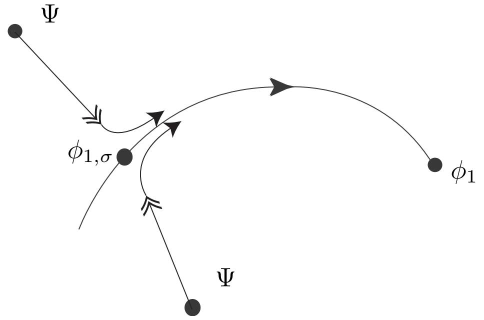


FIG. 1.1.

for a slowly varying real function θ. We claim that such a function $\sigma(\Psi)$ exists. When Ψ solves (Σ), we have

$$\begin{aligned}\frac{d\mathcal{V}}{dt} &= -2v(\Psi)\Im\Big(\sum_{k=1}^{N} a_k\left\langle x\Psi\ ,\ \phi_{k,\sigma(\Psi)}\right\rangle\left\langle\phi_{k,\sigma(\Psi)}\ ,\ \Psi\right\rangle\Big)\\ &\quad-\frac{d\sigma(\Psi)}{dt}2\Re\Big(\sum_{k=1}^{N} a_k\langle\Psi,\phi_{k,\sigma(\Psi)}\rangle\Big\langle\frac{d\phi_{k,\sigma(\Psi)}}{d\sigma},\Psi\Big\rangle\Big),\end{aligned}$$

where $\Re$ denotes the real part of a complex number, $(a_k)_{1\leqslant k\leqslant N}$ is defined by (1.8), and the notation $\frac{d\phi_{k,\sigma(\Psi)}}{d\sigma}$ means the derivative of the map $\sigma\mapsto\phi_{k,\sigma}$ taken at the point $\sigma=\sigma(\Psi)$. By definition of $\sigma(\Psi)$, we have

$$\frac{d\sigma(\Psi)}{dt}=\theta'(\mathcal{V}(\Psi))\frac{d\mathcal{V}}{dt}.$$

Thus, the feedback law $u(\Psi):=\sigma(\Psi)+v(\Psi)$, where

$$v(\Psi):=\varsigma\Im\Big(\sum_{k=1}^{N} a_k\left\langle x\Psi\ ,\ \phi_{k,\sigma(\Psi)}\right\rangle\left\langle\phi_{k,\sigma(\Psi)}\ ,\ \Psi\right\rangle\Big)$$

with $\varsigma>0$, ensures

$$\frac{d\mathcal{V}}{dt}=-2\varsigma\mu v(\Psi)^2,$$

where

$$\frac{1}{\mu}=1+2\theta'(\mathcal{V}(\Psi))\Re\Big(\sum_{k=1}^{N} a_k\langle\Psi,\phi_{k,\sigma(\Psi)}\rangle\Big\langle\frac{d\phi_{k,\sigma(\Psi)}}{d\sigma},\Psi\Big\rangle\Big)$$

is a positive constant, when $\|\theta'\|_{L^\infty}$ is small enough. Thus $t\mapsto\mathcal{V}(\Psi(t))$ is not increasing.

We claim that the solution of (1.1)–(1.3) with initial condition Ψ_0 and control $u=\sigma(\Psi)+v(\Psi)$ satisfies

$$\limsup_{t\to+\infty}\operatorname{dist}_{L^2}(\Psi(t),\mathcal{C}_1)^2\leqslant\epsilon. \tag{1.13}$$

Again, this will be proved by studying the $L^2(I,\mathbb{C})$-weak limits of $\Psi(t)$ when $t\to+\infty$.

1.4. Structure of the article. The rest of the paper is organized as follows.

Section 2 is dedicated to the proof of Theorem 1.2 when $\sigma \neq 0$. We derive this theorem as a consequence of a stronger result stated in Theorem 2.1.

This theorem and a straightforward corollary (Corollary 2.2), leading to Theorem 1.2 in the case $\sigma \neq 0$, will be stated in subsection 2.1. Subsection 2.2 is dedicated to some preliminary study needed for the proofs of Theorem 2.1 and Corollary 2.2. The proofs will be detailed in subsection 2.3.

Section 3 is devoted to the proof of Theorem 1.2 in the case $\sigma = 0$. Again, this theorem will be derived as a consequence of a stronger result stated in Theorem 3.2.

In subsection 3.1, we state a proposition (Proposition 3.1) ensuring the existence of the implicit function $\sigma = \sigma(\Psi)$. Then we state Theorem 3.2 and a straightforward corollary (Corollary 3.3), leading to Theorem 1.2 in the case $\sigma = 0$. A preliminary study, in preparation of the proofs of Theorem 3.2 and Corollary 3.3, will be performed in subsection 3.2. The proofs will be detailed in subsection 3.3.

Finally, in section 4, we provide some numerical simulations to check out the performance of the control design on a rather hard test case.

2. Stabilization of $\mathcal{C}_{1,\sigma}$ with $\sigma \neq 0$.

2.1. Main result.

The main result of section 2 is the following theorem.

THEOREM 2.1. *Let $N \in \mathbb{N}^*$. There exists $\sigma^\sharp = \sigma^\sharp(N) > 0$ such that, for every $\sigma \in (-\sigma^\sharp, \sigma^\sharp) - \{0\}$, $\gamma \in (0,1)$, $\epsilon > 0$, and $\Psi_0 \in \mathbb{S}$ verifying*

$$\sum_{k=N+1}^{\infty} |\langle \Psi_0, \phi_{k,\sigma}\rangle|^2 < \frac{\epsilon\gamma^2}{1-\epsilon} \quad \textit{and} \quad |\langle \Psi_0, \phi_{1,\sigma}\rangle| \geqslant \gamma, \tag{2.1}$$

the Cauchy problem (1.1)–(1.3) *with $u(t) = \sigma + v_{\sigma,N,\epsilon}(\Psi(t))$,*

$$v_{\sigma,N,\epsilon}(\Psi) := -\Im\left((1-\epsilon)\sum_{k=1}^{N}\langle x\Psi, \phi_{k,\sigma}\rangle\overline{\langle \Psi, \phi_{k,\sigma}\rangle} + \epsilon\langle x\Psi, \phi_{1,\sigma}\rangle\overline{\langle \Psi, \phi_{1,\sigma}\rangle}\right), \tag{2.2}$$

has a unique weak solution Ψ. Moreover, this solution satisfies

$$\liminf_{t\to+\infty} |\langle \Psi(t), \phi_{1,\sigma}\rangle|^2 \geqslant 1-\epsilon. \tag{2.3}$$

Theorem 2.1 provides an almost global practical stabilization. Indeed, any initial condition $\Psi_0 \in \mathbb{S}$ such that $\langle \Psi_0, \phi_{1,\sigma}\rangle \neq 0$ can be approximately moved to $\mathcal{C}_{1,\sigma}$. Notice that the regularity assumption $\Psi_0 \in H^s_{(\sigma)}(I,\mathbb{C})$ stated in Theorem 1.2 is not necessary for this purpose. Indeed, the feedback law depends on the initial state through the choice of the cut-off dimension N.

The following corollary states that the quantity N appearing in the feedback law may be uniform when Ψ_0 is in a given bounded subset of $H^s_{(\sigma)}(I,\mathbb{C})$.

COROLLARY 2.2. *Let $s > 0$, $\epsilon > 0$, $\Gamma > 0$, and $\gamma \in (0,1)$. There exist $\sigma^{**} = \sigma^{**}(\Gamma, s, \epsilon, \gamma) > 0$ and $N = N(\Gamma, s, \epsilon, \gamma) \in \mathbb{N}^*$ such that, for every $\sigma \in (-\sigma^{**}, \sigma^{**}) - \{0\}$, and $\Psi_0 \in H^s_{(\sigma)}(I,\mathbb{C}) \cap \mathbb{S}$ verifying*

$$\|\Psi_0\|_{H^s_{(\sigma)}} \leqslant \Gamma \textit{ and } |\langle \Psi_0, \phi_{1,\sigma}\rangle| \geqslant \gamma, \tag{2.4}$$

the Cauchy problem (1.1)–(1.3) *with $u = \sigma + v_{\sigma,N,\epsilon}(\Psi)$ has a unique weak solution Ψ. Moreover, this solution satisfies* (2.3).

Remark 2. Theorem 1.2 in the case $\sigma \neq 0$ is a direct consequence of the previous corollary. The feedback law mentioned in Theorem 1.2 is explicitly given via Corollary 2.2 and Theorem 2.1.

Notice that, in the particular case $\sigma \neq 0$, Corollary 2.2 is slightly more general than Theorem 1.2. In fact, the assumption "$\Psi_0 \in H^2 \cap H_0^1(I, \mathbb{C})$" is not needed as we deal with weak solutions instead of strong ones. Trivially, this solution will be a strong solution for $\Psi_0 \in H^2 \cap H_0^1(I, \mathbb{C})$.

For $\sigma = 0$, this will no longer be the case: we will need solutions in $C^1(\mathbb{R}, L^2)$ (for which the assumption $\Psi_0 \in H^2 \cap H_0^1(I, \mathbb{C})$ is needed; see Proposition 1.1).

2.2. Preliminaries.

This section is devoted to the preliminary results that will be applied in the proof of Theorem 2.1.

2.2.1. Eigenvalues and eigenvectors of A_σ.

PROPOSITION 2.3. *For every $k \in \mathbb{N}^*$, the eigenvalue $\sigma \mapsto \lambda_{k,\sigma} \in \mathbb{R}$ and the eigenstate $\sigma \mapsto \phi_{k,\sigma} \in (H^2 \cap H_0^1)(I, \mathbb{C})$ are analytic functions of $\sigma \in \mathbb{R}$ around $\sigma = 0$, and the expansion $\lambda_{k,\sigma} = \lambda_k + \sigma^2 \lambda_k^{(2)} + o(\sigma^2)$ holds with*

$$\lambda_k^{(2)} = \frac{1}{24\pi^2 k^2} - \frac{5}{8\pi^4 k^4}. \tag{2.5}$$

There exist $\sigma^* > 0$, $C^* > 0$ such that, for every $\sigma_0, \sigma_1 \in (-\sigma^*, \sigma^*) - \{0\}$, for every $k \in \mathbb{N}^*$,

$$\langle x\phi_{1,\sigma_0}, \phi_{k,\sigma_0} \rangle \neq 0, \tag{2.6}$$

$$|\lambda_{k,\sigma_0} - \lambda_k| \leqslant \frac{C^* \sigma^2}{k}, \tag{2.7}$$

$$\left\| \frac{d\phi_{k,\sigma_0}}{d\sigma} \right\|_{L^2} \leqslant \frac{C^*}{k}, \tag{2.8}$$

$$\left\| \frac{d\phi_{k,\sigma_0}}{d\sigma} \right\|_{H_0^1} \leqslant C^*, \tag{2.9}$$

$$\|\phi_{k,\sigma_0} - \phi_{k,\sigma_1}\|_{L^2} \leqslant \frac{C^* |\sigma_0 - \sigma_1|}{k}. \tag{2.10}$$

In the previous proposition, the notation $\frac{d\phi_{k,\sigma_0}}{d\sigma}$ means the derivative of the map $\sigma \mapsto \phi_{k\sigma}$ taken at the point $\sigma = \sigma_0$. In the same way, we will use the notation $\frac{d\lambda_{k,\sigma_0}}{d\sigma}$ for the derivative of the map $\sigma \mapsto \lambda_{k,\sigma}$ at $\sigma = \sigma_0$.

Proof of Proposition 2.3. We consider the family of self-adjoint operators $A_\sigma = A - \sigma x$ in the space $(H^2 \cap H_0^1)(I, \mathbb{C})$. In this Banach space, the operator x (as a multiplication operator) is relatively bounded with respect to A with relative bound 0 (in the sense of [15, p. 190]). Therefore A_σ is a self-adjoint holomorphic family of type (A) (see [15, p. 375]). Thus the eigenvalues and the eigenstates of A_σ are holomorphic functions of σ.

Thanks to the Rayleigh–Schrödinger perturbation theory, we compute the first terms of the expansions

$$\lambda_{k,\sigma} = \lambda_k + \sigma\lambda_k^{(1)} + \sigma^2\lambda_k^{(2)} + \cdots, \quad \phi_{k,\sigma} = \phi_k + \sigma\phi_k^{(1)} + \sigma^2\phi_k^{(2)} + \cdots.$$

Considering the first and second order terms of the equalities $A_\sigma \phi_{k,\sigma} = \lambda_{k,\sigma}\phi_{k,\sigma}$, $\|\phi_{k,\sigma}\|^2_{L^2} = 1$, we get

$$-\frac{1}{2}\frac{d^2}{dx^2}\phi_k^{(1)} - x\phi_k = \lambda_k \phi_k^{(1)} + \lambda_k^{(1)}\phi_k, \qquad \langle \phi_k^{(1)}, \phi_k\rangle = 0, \tag{2.11}$$

$$-\frac{1}{2}\frac{d^2}{dx^2}\phi_k^{(2)} - x\phi_k^{(1)} = \lambda_k \phi_k^{(2)} + \lambda_k^{(1)}\phi_k^{(1)} + \lambda_k^{(2)}\phi_k, \quad 2\Re\langle \phi_k^{(2)}, \phi_k\rangle + \|\phi_k^{(1)}\|^2_{L^2} = 0. \tag{2.12}$$

Taking the Hermitian product of the first equality of (2.11) with ϕ_k and applying the parity properties of ϕ_k, we get $\lambda_k^{(1)} = 0$. Considering the Hermitian product of the first equality of (2.11) with ϕ_j, we get

$$\phi_k^{(1)} = \sum_{j\in\mathbb{N}^*, P(j)\neq P(k)} \frac{\langle x\phi_j, \phi_k\rangle}{\lambda_j - \lambda_k}\phi_j, \tag{2.13}$$

where the sum is taken over $j \in \mathbb{N}^*$ such that the parity of j is different from the parity of k. Taking the Hermitian product of the first equality of (2.12) with ϕ_k, we get $\lambda_k^{(2)} = -\langle x\phi_k^{(1)}, \phi_k\rangle$. Using (2.13) and the explicit expression of $\langle x\phi_k, \phi_j\rangle$ computed thanks to (1.5), we get

$$\lambda_k^{(2)} = \frac{2^7}{\pi^6} \sum_{j\in\mathbb{N}^*, P(j)\neq P(k)} \frac{k^2 j^2}{(k^2 - j^2)^5}. \tag{2.14}$$

In order to simplify the above sum, we decompose the fraction

$$F(X) := \frac{X^2}{(X-q)^5 (X+q)^5}$$

in the form

$$\begin{aligned} F(X) = \frac{1}{2^5 q^3}\left(\frac{1}{(X-q)^5} - \frac{1}{(X+q)^5}\right) - \frac{1}{2^6 q^4}\left(\frac{1}{(X-q)^4} + \frac{1}{(X+q)^4}\right) \\ - \frac{1}{2^7 q^5}\left(\frac{1}{(X-q)^3} - \frac{1}{(X+q)^3}\right) + \frac{5}{2^8 q^6}\left(\frac{1}{(X-q)^2} + \frac{1}{(X+q)^2}\right) \\ - \frac{5}{2^8 q^7}\left(\frac{1}{(X-q)} - \frac{1}{(X+q)}\right). \end{aligned}$$

Inserting this relation in the sum (2.14) and simplifying, we find

$$\lambda_k^{(2)} = \frac{1}{\pi^6}\left(\frac{5}{2k^5}S_k^1 - \frac{5}{2k^4}S_k^2 + \frac{1}{k^3}S_k^3 + \frac{2}{k^2}S_k^4 - \frac{4}{k}S_k^5\right), \tag{2.15}$$

where

$$S_k^a := \sum_{j\in\mathbb{N}^*, P(j)\neq P(k)} \left(\frac{1}{(j-k)^a} + \frac{(-1)^a}{(j+k)^a}\right) \text{ for } a = 1, \dots, 5.$$

We apply now the following well-known relations for the Riemann ζ-function:

$$\zeta(2) = \sum_{j=1}^{\infty} \frac{1}{j^2} = \frac{\pi^2}{6} \qquad \text{and} \qquad \zeta(4) = \sum_{j=1}^{\infty} \frac{1}{j^4} = \frac{\pi^4}{90}.$$

These relations imply

$$\sum_{k=-\infty}^{\infty} \frac{1}{(2j+1)^2} = \frac{\pi^2}{4} \qquad \text{and} \qquad \sum_{k=-\infty}^{\infty} \frac{1}{(2j+1)^4} = \frac{\pi^4}{48};$$

thus

$$S_k^a = \begin{cases} \frac{1}{k^a} & \text{when } k \text{ is odd} \\ 0 & \text{when } k \text{ is even} \end{cases} \quad \text{for } a = 1, 3, 5,$$

$$S_k^2 = \begin{cases} \frac{\pi^2}{4} - \frac{1}{k^2} & \text{when } k \text{ is odd,} \\ \frac{\pi^2}{4} & \text{when } k \text{ is even,} \end{cases}$$

$$S_k^4 = \begin{cases} \frac{\pi^4}{48} - \frac{1}{k^4} & \text{when } k \text{ is odd,} \\ \frac{\pi^4}{48} & \text{when } k \text{ is even.} \end{cases}$$

Inserting this in (2.15), we get (2.5).

The relation (2.6) is proved in [5, Proposition 1]. The bound (2.7) is given in [15, Chapter 17, Example 2.14, Chapter 2, Problem 3.7]. Inequality (2.8) is proved in [5, Proposition 42]. The bound (2.9) is a consequence of (2.8). Indeed, considering the Hermitian product in $L^2(I, \mathbb{C})$ of $\frac{d\phi_{k,\sigma_0}}{d\sigma}$ with the equation

$$A_{\sigma_0} \frac{d\phi_{k,\sigma_0}}{d\sigma} - x\phi_{k,\sigma_0} = \lambda_{k,\sigma_0} \frac{d\phi_{k,\sigma_0}}{d\sigma} + \frac{d\lambda_{k,\sigma_0}}{d\sigma} \phi_{k,\sigma_0},$$

and using (2.8) together with the orthogonality between ϕ_{k,σ_0} and $\frac{d\phi_{k,\sigma_0}}{d\sigma}$ (which is a consequence of $\|\phi_{k,\sigma}\|_{L^2}^2 \equiv 1$), we get

$$\left\| \frac{d\phi_{k,\sigma_0}}{d\sigma} \right\|_{H_0^1}^2 \leqslant |\sigma_0| \left(\frac{C^*}{k} \right)^2 + \frac{C^*}{k} + \left(\frac{\pi^2 k^2}{2} + C^* \sigma_0^2 \right) \left(\frac{C^*}{k} \right)^2,$$

which gives (2.9). Finally, (2.10) is a consequence of (2.8). □

Proposition 2.4. *Let $N \in \mathbb{N}^*$. There exists $\sigma^\sharp = \sigma^\sharp(N) > 0$ such that, for every $\sigma \in (-\sigma^\sharp, \sigma^\sharp) - \{0\}$, $j_2, k_2 \in \mathbb{N}^*$, and $j_1, k_1 \in \{1, \ldots, N\}$, verifying $j_1 \neq j_2$ and $k_1 \neq k_2$,*

$$\lambda_{k_1,\sigma} - \lambda_{k_2,\sigma} = \lambda_{j_1,\sigma} - \lambda_{j_2,\sigma} \tag{2.16}$$

implies $(j_1, j_2) = (k_1, k_2)$.

Proof of Proposition 2.4. Let C^* be as in Proposition 2.3 and $\sigma \in (-\sigma_0^\sharp, \sigma_0^\sharp)$ where

$$\sigma_0^\sharp := \frac{\pi}{4\sqrt{C^*}}. \tag{2.17}$$

First, we prove (2.16) to be impossible when $j_2 \neq k_2$ and

$$\max\{j_2, k_2\} > \frac{N^2+1}{2}. \tag{2.18}$$

We argue by contradiction. Let us assume the existence of $j_2, k_2 \in \mathbb{N}^*$, $j_1, k_1 \in \{1, \ldots, N\}$, with $j_1 \neq j_2$, $k_1 \neq k_2$, $j_2 \neq k_2$, such that (2.18) and (2.16) hold. Without loss of generality, we may assume that $\max\{j_2,k_2\} = j_2 > \frac{N^2+1}{2}$. Using (2.7), we get

$$\begin{aligned}\lambda_{j_2,\sigma} - \lambda_{k_2,\sigma} &\geqslant \frac{\pi^2}{2}(j_2^2 - k_2^2) - 2C^*\sigma^2\\ &\geqslant \frac{\pi^2}{2}(j_2^2 - (j_2-1)^2) - 2C^*\sigma^2\\ &\geqslant \frac{\pi^2}{2}(2j_2 - 1) - 2C^*\sigma^2,\\ \lambda_{j_1,\sigma} - \lambda_{k_1,\sigma} &\leqslant \frac{\pi^2}{2}(N^2-1) + 2C^*\sigma^2.\end{aligned}$$

Using the equality of the left-hand sides of these inequalities, together with (2.17), we get

$$j_2 \leqslant \frac{N^2}{2} + \frac{8C^*\sigma_0^{\sharp 2}}{\pi^2} \leqslant \frac{N^2+1}{2},$$

which is a contradiction.

Therefore, it is sufficient to prove Proposition 2.4 for $j_2, k_2 \in \{1,\dots,[(N^2+1)/2]\}$. Moreover, it is sufficient to prove that, for every $j_1, k_1 \in \{1,\dots,N\}$ and $j_2,k_2 \in \{1,\dots,[(N^2+1)/2]\}$, with $j_1 \neq j_2$, $k_1 \neq k_2$, $(j_1,j_2) \neq (k_1,k_2)$, there exists $\sigma^\sharp_{j_1,k_1,j_2,k_2} \in (0,\sigma_0^\sharp)$ such that, for every $\sigma \in (-\sigma^\sharp_{j_1,k_1,j_2,k_2}, \sigma^\sharp_{j_1,k_1,j_2,k_2})$, (2.16) does not hold. Indeed, then, the following choice of $\sigma^\sharp(N)$ concludes the proof of Proposition 2.4:

$$\begin{aligned}\sigma^\sharp(N) := \min\{\sigma^\sharp_{j_1,k_1,j_2,k_2};\quad & j_1,k_1 \in \{1,\dots,N\}, j_2,k_2 \in \{1,\dots,(N^2+1)/2\},\\ & (j_1,j_2) \neq (k_1,k_2), j_1 \neq j_2, k_1 \neq k_2\}.\end{aligned}$$

Let $j_1,k_1 \in \{1,\dots,N\}$, $j_2,k_2 \in \{1,\dots,(N^2+1)/2\}$ be such that $j_1 \neq j_2$, $k_1 \neq k_2$, $(j_1,j_2) \neq (k_1,k_2)$. We argue by contradiction. Let us assume that, for every $\sigma_1^\sharp > 0$, there exists $\sigma \in (-\sigma_1^\sharp, \sigma_1^\sharp)$ such that (2.16) holds. Using the analyticity of both sides in (2.16) with respect to σ, at $\sigma = 0$, this assumption implies that

$$\lambda_{k_1}^{(2)} - \lambda_{k_2}^{(2)} = \lambda_{j_1}^{(2)} - \lambda_{j_2}^{(2)}.$$

Using (2.5) together with a rationality argument, we get

$$\frac{1}{k_1^2} - \frac{1}{k_2^2} = \frac{1}{j_1^2} - \frac{1}{j_2^2}, \quad \frac{1}{k_1^4} - \frac{1}{k_2^4} = \frac{1}{j_1^4} - \frac{1}{j_2^4}.$$

Since $k_1 \neq k_2$ and $j_1 \neq j_2$, we deduce from the previous equalities that

$$\frac{1}{k_1^2} - \frac{1}{k_2^2} = \frac{1}{j_1^2} - \frac{1}{j_2^2}, \quad \frac{1}{k_1^2} + \frac{1}{k_2^2} = \frac{1}{j_1^2} + \frac{1}{j_2^2}.$$

Therefore $k_1 = j_1$ and $k_2 = j_2$, which is a contradiction. □

2.2.2. Solutions of the Cauchy problem.

PROPOSITION 2.5. *Let $\sigma \in \mathbb{R}$, $N \in \mathbb{N}^*$, $\epsilon > 0$. For every $\Psi_0 \in \mathbb{S}$, there exists a unique weak solution Ψ of* (1.1)–(1.3) *with $u(t) = \sigma + v_{\sigma,N,\epsilon}(\Psi(t))$, i.e., $\Psi \in C^0(\mathbb{R},\mathbb{S}) \cap C^1(\mathbb{R}, H^{-2}_{(0)}(I,\mathbb{C}))$,* (1.1) *holds in $H^{-2}_{(0)}(I,\mathbb{C})$ for every $t \in \mathbb{R}$, and* (1.2) *holds in $\mathbb{S}$.*

Proof of Proposition 2.5. Let $\sigma \in \mathbb{R}$, $N \in \mathbb{N}$, $\epsilon > 0$, $\Psi_0 \in \mathbb{S}$, and $T > 0$ be such that

$$TNe^{NT} < 1. \tag{2.19}$$

In order to build solutions on $[0,T]$, we apply the Banach fixed point theorem to the map

$$\begin{array}{rccc} \Theta: & C^0([0,T],\mathbb{S}) & \to & C^0([0,T],\mathbb{S}) \\ & \xi & \mapsto & \Psi, \end{array}$$

where Ψ is the solution of (1.1)–(1.3) with $u(t) = \sigma + v_{\sigma,N,\epsilon}(\xi(t))$.

The map Θ is well defined and maps $C^0([0,T],\mathbb{S})$ into itself. Indeed, when $\xi \in C^0([0,T],\mathbb{S})$, $u : t \mapsto \sigma + v_{\sigma,N,\epsilon}(\xi(t))$ is continuous and thus Proposition 1.1 ensures the existence of a unique weak solution Ψ. Notice that the map Θ takes values in $C^0([0,T],\mathbb{S}) \cap C^1([0,T],H^{-2}_{(0)})$.

Let us prove that Θ is a contraction of $C^0([0,T],\mathbb{S})$. Let $\xi_j \in C^0([0,T],\mathbb{S})$, $v_j := v_{\sigma,N,\epsilon}(\xi_j)$, $\Psi_j := \Theta(\xi_j)$ for $j = 1,2$ and $\Delta := \Psi_1 - \Psi_2$. We have

$$\Delta(t) = i\int_0^t e^{-iA_\sigma(t-s)}[v_1 x\Delta(s) + (v_1 - v_2)x\Psi_2(s)]ds.$$

Thanks to (2.2), we have $\|v_j\|_{L^\infty(0,T)} \leqslant N$ for $j = 1,2$ and $\|v_1 - v_2\|_{L^\infty(0,T)} \leqslant 2N\|\xi_1 - \xi_2\|_{C^0([0,T],L^2)}$. Thus

$$\|\Delta(t)\|_{L^2} \leqslant \int_0^t N\|\Delta(s)\|_{L^2} + N\|\xi_1 - \xi_2\|_{C^0([0,T],L^2)}ds. \tag{2.20}$$

Therefore, the Gronwall lemma implies

$$\|\Delta(t)\|_{C^0([0,T],L^2)} \leqslant \|\xi_1 - \xi_2\|_{C^0([0,T],L^2)} NTe^{NT},$$

and so (2.19) ensures that Θ is a contraction of the Banach space $C^0([0,T],\mathbb{S})$. Therefore, there exists a fixed point $\Psi \in C^0([0,T],\mathbb{S})$ such that $\Theta(\Psi) = \Psi$. Since Θ takes values in $C^0([0,T],\mathbb{S}) \cap C^1([0,T],H^{-2}_{(0)})$, necessarily Ψ belongs to this space, and thus it is a weak solution of (1.1)–(1.3) on $[0,T]$.

Finally, we have introduced a time $T > 0$ and, for every $\Psi_0 \in \mathbb{S}$, we have built a weak solution $\Psi \in C^0([0,T],\mathbb{S})$ of (1.1)–(1.3) on $[0,T]$. Thus, for a given initial condition $\Psi_0 \in \mathbb{S}$, we can apply this result on $[0,T]$, $[T,2T]$, $[2T,3T]$, etc. This proves the existence and uniqueness of a global solution for the closed-loop system. □

PROPOSITION 2.6. *Let $\sigma > 0$, $N \in \mathbb{N}$, $\epsilon > 0$, $(\Psi_0^n)_{n\in\mathbb{N}}$ be a sequence of $\mathbb{S}$, and let $\Psi_0^\infty \in L^2$ with $\|\Psi_0^\infty\|_{L^2} \leqslant 1$ be such that*

$$\lim_{n\to+\infty} \Psi_0^n = \Psi_0^\infty \textit{ strongly in } H^{-1}(I,\mathbb{C}).$$

Let Ψ^n (resp., Ψ^∞) be the weak solution of (1.1)–(1.3) *with $u(t) = \sigma + v_{\sigma,N,\epsilon}(\Psi^n)$ (resp., with $u(t) = \sigma + v_{\sigma,N,\epsilon}(\Psi^\infty(t))$). Then, for every $\tau > 0$,*

$$\lim_{n\to+\infty} \Psi^n(\tau) = \Psi^\infty(\tau) \textit{ strongly in } H^{-1}(I,\mathbb{C}).$$

Proof of Proposition 2.6. Let us recall that the space $H^{-1}(I,\mathbb{C})$ (dual space of $H^1_0(I,\mathbb{C})$ for the $L^2(I,\mathbb{C})$-Hermitian product) coincides with $H^{-1}_{(0)}(I,\mathbb{C})$ and that $\sqrt{2}\|.\|_{H^{-1}} = \|.\|_{H^{-1}_{(0)}}$ (because $\|.\|_{H^1_0} = \sqrt{2}\|.\|_{H^1_{(0)}}$). We introduce $\mathcal{C} > 0$ such that

$$\|x\varphi\|_{H^{-1}} \leqslant \mathcal{C}\|\varphi\|_{H^{-1}} \quad \forall\varphi \in H^{-1}(I,\mathbb{C}). \tag{2.21}$$

Such a constant does exist. Indeed, for every $\xi \in H^1_0(I,\mathbb{C})$, $x\xi \in H^1_0(I,\mathbb{C})$, and

$$\|x\xi\|_{H^1_0} = \left(\int_I |x\xi' + \xi|^2 dx\right)^{1/2} \leqslant \|\xi'\|_{L^2}(1 + C_P),$$

where C_P is the Poincaré constant on I. Thus, for $\varphi \in H^{-1}(I,\mathbb{C})$, we have

$$\begin{aligned}\|x\varphi\|_{H^{-1}(I,\mathbb{C})} &= \sup\left\{\langle x\varphi, \xi\rangle; \xi \in H^1_0(I,\mathbb{C}), \|\xi\|_{H^1_0} = 1\right\} \\ &\leqslant \sup\left\{\|\varphi\|_{H^{-1}}\|x\xi\|_{H^1_0}; \xi \in H^1_0(I,\mathbb{C}), \|\xi\|_{H^1_0} = 1\right\} \\ &\leqslant (1 + C_P)\|\xi\|_{H^{-1}}.\end{aligned}$$

In order to simplify the notation, in this proof we write $v(\Psi)$ instead of $v_{\sigma,N,\epsilon}(\Psi)$. We have

$$\begin{aligned}(\Psi^n - \Psi^\infty)(t) &= e^{-iAt}(\Psi^n_0 - \Psi^\infty_0) + i\int_0^t e^{-iA(t-s)}\sigma x(\Psi^n - \Psi^\infty)(s)ds \\ &\quad + i\int_0^t e^{-iA(t-s)}[v(\Psi^n(s)) - v(\Psi^\infty(s))]x\Psi^n(s)ds \\ &\quad + i\int_0^t e^{-iA(t-s)}v(\Psi^\infty(s))x[\Psi^n(s) - \Psi^\infty(s)]ds.\end{aligned}$$

Using (2.2), $\|\Psi^n(s)\|_{L^2} = 1$, $\|\Psi^\infty(s)\|_{L^2} \leqslant 1$ and the fact that $\phi_{k,\sigma}, x\phi_{k,\sigma} \in H^1_0(I,\mathbb{C})$ for $k = 1,\dots,N$, we get

$$|v(\Psi^n(s)) - v(\Psi^\infty(s))| \leqslant 2N\mathcal{C}C_\sigma(N)\|(\Psi^n - \Psi^\infty)(s)\|_{H^{-1}}, \tag{2.22}$$

where $C_\sigma(N) := \sup\{\|\phi_{k,\sigma}\|_{H^1_0(I,\mathbb{C})}; k \in \{1,\dots,N\}\}$. The semigroup e^{-iAt} preserves the H^{-1}-norm, and thus, using $|v(\Psi^\infty(s))| \leqslant N$ and (2.22), we get

$$\begin{aligned}\|(\Psi^n - \Psi^\infty)(t)\|_{H^{-1}} &\leqslant \|\Psi^n_0 - \Psi^\infty_0\|_{H^{-1}} \\ &\quad + \mathcal{C}\textstyle\int_0^t(|\sigma| + 2NC_\sigma(N) + N)\|\Psi^n(s) - \Psi^\infty(s)\|_{H^{-1}}ds.\end{aligned}$$

We conclude thanks to the Gronwall lemma. □

2.3. Proofs of Theorem 2.1 and Corollary 2.2.

Proof of Theorem 2.1. Let $N \in \mathbb{N}^*$. Let $\sigma^* > 0$ be as in Proposition 2.3 and $\sigma^\sharp = \sigma^\sharp(N)$ be as in Proposition 2.4. Let $\sigma^{**} := \min\{\sigma^*, \sigma^\sharp\}$.

Let $\sigma \in (-\sigma^{**}, \sigma^{**}) - \{0\}$, $\gamma \in (0,1)$, $\epsilon > 0$, $\Psi_0 \in \mathbb{S}$ with (2.1) and let Ψ be the weak solution of (1.1)–(1.3) with $u(t) = \sigma + v_{\sigma,N,\epsilon}(\Psi(t))$ given by Proposition 2.5. For $\varphi \in L^2(I,\mathbb{C})$, we define

$$\mathcal{V}_{\sigma,N,\epsilon}(\varphi) := 1 - |\langle\varphi, \phi_{1,\sigma}\rangle|^2 - (1-\epsilon)\sum_{k=2}^N |\langle\varphi, \phi_{k,\sigma}\rangle|^2. \tag{2.23}$$

Since $\Psi \in C^1(\mathbb{R}, H^{-2}_{(0)}(I,\mathbb{C}))$ and $\phi_{k,\sigma} \in H^2_{(0)}(I,\mathbb{C})$, $t \mapsto \mathcal{V}_{N,\sigma,\epsilon}(\Psi(t))$ is C^1. Using (1.1), integration by parts, and $a_1 := 1$, $a_k := 1 - \epsilon$ when $k \geqslant 2$, we get

$$\begin{aligned}\frac{d}{dt}\mathcal{V}_{\sigma,N,\epsilon}(\Psi) &= -2\Re\left(\sum_{k=1}^N a_k\langle -iA_\sigma\Psi + iv_{\sigma,N,\epsilon}(\Psi)x\Psi, \phi_{k,\sigma}\rangle\overline{\langle\Psi, \phi_{k,\sigma}\rangle}\right), \\ &= -2v_{\sigma,N,\epsilon}(\Psi(t))^2.\end{aligned} \tag{2.24}$$

Thus, $t \mapsto \mathcal{V}_{\sigma,N,\epsilon}(\Psi(t))$ is a nonincreasing function. There exists $\alpha \in [0, \mathcal{V}_{\sigma,N,\epsilon}(\Psi_0)]$ such that $\mathcal{V}_{\sigma,N,\epsilon}(\Psi(t)) \to \alpha$ when $t \to +\infty$. Since $\Psi_0 \in \mathbb{S}$ and (2.1) holds we have

$$\begin{aligned}
\mathcal{V}_{\sigma,N,\epsilon}(\Psi_0) &= 1 - (1-\epsilon)\sum_{k=1}^{N} |\langle \Psi, \phi_{k,\sigma}\rangle|^2 - \epsilon |\langle \Psi, \phi_{1,\sigma}\rangle|^2 \\
&= 1 - (1-\epsilon)\left(1 - \sum_{k=N+1}^{\infty} |\langle \Psi, \phi_{k,\sigma}\rangle|^2\right) - \epsilon |\langle \Psi, \phi_{1,\sigma}\rangle|^2 \\
&< 1 - (1-\epsilon)\left(1 - \frac{\epsilon\gamma^2}{1-\epsilon}\right) - \epsilon\gamma^2 \\
&< \epsilon,
\end{aligned}$$

and thus $\alpha \in [0, \epsilon)$.

Let $(t_n)_{n\in\mathbb{N}}$ be an increasing sequence of positive real numbers such that $t_n \to +\infty$ when $n \to +\infty$. Since $\|\Psi(t_n)\|_{L^2} = 1$ for every $n \in \mathbb{N}$, there exists $\Psi_\infty \in L^2(I, \mathbb{C})$ such that, up to an extraction,

$$\Psi(t_n) \to \Psi_\infty \text{ weakly in } L^2(I,\mathbb{C}) \text{ and strongly in } H^{-1}(I,\mathbb{C}).$$

Let ξ be the solution of

$$\begin{cases} i\frac{\partial \xi}{\partial t} = A_\sigma \xi - v_{\sigma,N,\epsilon}(\xi(t))x\xi, x \in I, t \in (0,+\infty), \\ \xi(t, \pm 1/2) = 0, \\ \xi(0) = \Psi_\infty. \end{cases}$$

Thanks to Proposition 2.6, for every $\tau > 0$, $\Psi(t_n + \tau) \to \xi(\tau)$ strongly in $H^{-1}(I,\mathbb{C})$ when $n \to +\infty$. Thus $\mathcal{V}_{\sigma,N,\epsilon}(\Psi(t_n+\tau)) \to \mathcal{V}_{\sigma,N,\epsilon}(\xi(\tau))$ when $n \to +\infty$, because $\mathcal{V}_{\sigma,N,\epsilon}(.)$ is continuous for the L^2-weak topology. Therefore $\mathcal{V}_{\sigma,N,\epsilon}(\xi(\tau)) \equiv \alpha$. Furthermore, relation (2.24) holds when Ψ is replaced by ξ, and thus $v_{\sigma,N,\epsilon}(\xi(\tau)) \equiv 0$ and ξ solves

$$\begin{cases} i\frac{\partial \xi}{\partial t} = A_\sigma \xi, x \in I, t \in (0,+\infty), \\ \xi(t, \pm 1/2) = 0, \\ \xi(0) = \Psi_\infty. \end{cases}$$

Therefore, we have

$$\xi(\tau) = \sum_{k=1}^{\infty} \langle \Psi_\infty, \phi_{k,\sigma}\rangle \phi_{k,\sigma} e^{-i\lambda_{k,\sigma}\tau}.$$

The equality $v_{\sigma,N,\epsilon}(\xi) \equiv 0$, then, gives

(2.25)
$$\Im\left(\sum_{k=1}^{N} \sum_{j\in\mathbb{N}^*, j\neq k} a_k \langle \Psi_\infty, \phi_{j,\sigma}\rangle \langle x\phi_{j,\sigma}, \phi_{k,\sigma}\rangle \overline{\langle \Psi_\infty, \phi_{k,\sigma}\rangle} e^{i(\lambda_{k,\sigma} - \lambda_{j,\sigma})\tau}\right) \equiv 0.$$

Let $\omega_{(k_1,k_2)} := \lambda_{k_1,\sigma} - \lambda_{k_2,\sigma}$ for every $k_1, k_2 \in \mathbb{N}^*$ and $\mathcal{S} := \{(k_1,k_2); k_1 \in \{1,\dots,N\}, k_2 \in \mathbb{N}^*, k_1 \neq k_2\}$. Thanks to Proposition 2.4, all the frequencies ω_K for $K \in \mathcal{S}$ are different. Moreover, there exists a uniform gap $\delta > 0$ such that, for every $\omega, \tilde{\omega} \in \{\pm\omega_K; K \in \mathcal{S}\}$ with $\omega \neq \tilde{\omega}$, then $|\omega - \tilde{\omega}| \geqslant \delta$. Thus, for $T > 0$ large enough, there exists $C = C(T) > 0$ such that the Ingham inequality

$$\sum_{K\in\mathcal{S}} |a_K|^2 \leqslant C \int_0^T \Big| \sum_{K\in\mathcal{S}} a_K e^{i\omega_K t} \Big|^2 dt$$

holds for every $(a_K)_{K\in\mathcal{S}} \in l^2(\mathcal{S}, \mathbb{C})$ (see [16, Theorem 1.2.9]). Equality (2.25) implies, in particular,

$$\langle \Psi_\infty, \phi_{j,\sigma}\rangle \langle x\phi_{j,\sigma}, \phi_{1,\sigma}\rangle \overline{\langle \Psi_\infty, \phi_{1,\sigma}\rangle} = 0 \quad \forall j \geqslant 2.$$

Thanks to (2.6), we get

$$\langle \Psi_\infty, \phi_{j,\sigma}\rangle \overline{\langle \Psi_\infty, \phi_{1,\sigma}\rangle} = 0 \quad \forall j \geqslant 2. \tag{2.26}$$

Let us prove that

$$\langle \Psi_\infty, \phi_{1,\sigma}\rangle \neq 0. \tag{2.27}$$

Since $\|\Psi^\infty\|_{L^2} \leqslant 1$, we have

$$\begin{aligned}\mathcal{V}_{\sigma,N,\epsilon}(\Psi_\infty) &\geqslant 1 - |\langle \Psi^\infty, \phi_{1,\sigma}\rangle|^2 - (1-\epsilon)\sum_{k=2}^{\infty} |\langle \Psi^\infty, \phi_{k,\sigma}\rangle|^2 \\ &= 1 - |\langle \Psi^\infty, \phi_{1,\sigma}\rangle|^2 - (1-\epsilon)[\|\Psi^\infty\|_{L^2}^2 - |\langle \Psi^\infty, \phi_{1,\sigma}\rangle|^2] \\ &\geqslant \epsilon - \epsilon |\langle \Psi^\infty, \phi_{1,\sigma}\rangle|^2.\end{aligned}$$

Moreover, $\mathcal{V}_{\sigma,N,\epsilon}(\Psi_\infty) = \alpha < \epsilon$, and thus

$$\epsilon > \epsilon - \epsilon |\langle \Psi_\infty, \phi_{1,\sigma}|^2,$$

which gives (2.27). Therefore (2.26) justifies the existence of $\beta \in \mathbb{C}$ with $|\beta| \leqslant 1$ such that $\Psi_\infty = \beta\phi_{1,\sigma}$. Then $\epsilon > \alpha = \mathcal{V}_{N,\sigma,\epsilon}(\Psi_\infty) = 1 - |\beta|^2$, and thus $|\beta|^2 > 1 - \epsilon$. Finally, we have

$$\lim_{n\to+\infty} |\langle \Psi(t_n), \phi_{1,\sigma}\rangle|^2 = |\langle \Psi_\infty, \phi_{1,\sigma}\rangle|^2 = |\beta|^2 > 1 - \epsilon.$$

This holds for every sequence $(t_n)_{n\in\mathbb{N}}$, and thus (2.3) is proved. □

Proof of Corollary 2.2. Let $C^*, \sigma^* > 0$ be as in Proposition 2.3. There exists $N = N(\Gamma, s, \epsilon, \gamma) \in \mathbb{N}^*$ large enough so that

$$\frac{\Gamma^2}{\left(\lambda_{N+1} - \frac{C^*\sigma^{*2}}{N+1}\right)^s} \leqslant \frac{\epsilon\gamma^2}{1-\epsilon}. \tag{2.28}$$

Let $\sigma^{**} = \sigma^{**}(N)$ be as in Theorem 2.1 (notice that $\sigma^{**} \leqslant \sigma^*$) and $\sigma \in (-\sigma^{**}, \sigma^{**}) - \{0\}$. Let $\Psi_0 \in H^s_{(\sigma)}(I, \mathbb{C}) \cap \mathbb{S}$, verifying (2.4). In order to get the conclusion of Corollary 2.2, we prove that (2.1) holds, and we apply Theorem 2.1. Using (2.7), we get

$$\begin{aligned}\sum_{k=N+1}^{\infty} |\langle \Psi_0, \phi_{k,\sigma}\rangle|^2 &\leqslant \frac{1}{\lambda_{N+1,\sigma}^s} \sum_{k=N+1}^{\infty} \lambda_{k,\sigma}^s |\langle \Psi_0, \phi_{k,\sigma}\rangle|^2 \\ &\leqslant \frac{1}{\lambda_{N+1,\sigma}^s} \sum_{k=1}^{\infty} \lambda_{k,\sigma}^s |\langle \Psi_0, \phi_{k,\sigma}\rangle|^2 \\ &\leqslant \frac{\Gamma^2}{\left(\lambda_{N+1} - \frac{C^*\sigma^2}{N+1}\right)^s}.\end{aligned}$$

Thus (2.28) implies (2.1). □

3. Stabilization of $\mathcal{C}_1$. Throughout this section, the constants C^*, σ^* are as in Proposition 2.3.

3.1. Main result. First, let us state the existence of an implicit function $\sigma(\Psi)$ that will be used in the feedback law. When X is a normed space, $a \in X$ and $r > 0$, we use the notation $B_X(a,r) := \{y \in X; \|y-a\|_X < r\}$.

PROPOSITION 3.1. *Let $N \in \mathbb{N}^*$, $\epsilon > 0$, and $\theta \in C^\infty(\mathbb{R}_+, [0,\sigma^*])$ be such that*

$$\theta(0) = 0, \quad \theta(s) > 0 \quad \forall s > 0, \quad \|\theta'\|_{L^\infty} \leqslant \frac{1}{36NC^*}. \tag{3.1}$$

There exists a unique $\sigma \in C^\infty(B_{L^2}(0,2), [0, \|\theta\|_{L^\infty}])$ such that

$$\sigma(\psi) = \theta(\mathcal{V}_{\sigma(\psi),N,\epsilon}(\psi)) \qquad \forall \psi \in B_{L^2}(0,2),$$

where $\mathcal{V}_{\sigma,N,\epsilon}$ is defined by (2.23).

The proof of this proposition is done in [7]. For the sake of completeness, we repeat it in the appendix. The main result of this section is the following.

THEOREM 3.2. *Let $N \in \mathbb{N}^*$, $\gamma \in (0,1)$, $\epsilon > 0$, $\theta \in C^\infty(\mathbb{R}_+, [0,\sigma^*])$ verifying* (3.1),

$$\|\theta\|_{L^\infty} \leqslant \min\left\{ \frac{1}{C^*}\left(\frac{\epsilon\gamma^2 N}{32(1-\epsilon/2)}\right)^{1/2}, \frac{\gamma}{2C^*}, \sigma^\sharp(N), \frac{1}{C^*}(\sqrt{1-\epsilon/2} - \sqrt{1-\epsilon}) \right\}, \tag{3.2}$$

and

$$\|\theta'\|_{L^\infty} < \frac{1}{3(1+NC^*)}. \tag{3.3}$$

Let $\sigma \in C^\infty(B_{L^2}(0,2), [0, \|\theta\|_{L^\infty}])$ be as in Proposition 3.1. *For every $\Psi_0 \in \mathbb{S} \cap (H^2 \cap H_0^1)(I,\mathbb{C})$ with*

$$\sum_{k=N+1}^{\infty} |\langle \Psi_0, \phi_k\rangle|^2 < \frac{\epsilon\gamma^2}{32(1-\epsilon/2)} \quad \text{and} \quad |\langle \Psi_0, \phi_1\rangle| \geqslant \gamma, \tag{3.4}$$

the Cauchy problem (1.1)–(1.3) *with $u(t) = \sigma(\Psi(t)) + v_{\sigma(\psi(t)),N,\epsilon}(\Psi(t))$ has a unique strong solution ψ. Moreover this solution satisfies*

$$\liminf_{t\to+\infty} |\langle \Psi(t), \phi_1\rangle|^2 \geqslant 1 - \epsilon. \tag{3.5}$$

The following corollary states that the quantity N appearing in the feedback law may be uniform in a fixed bounded subset of H^s for $s > 0$.

COROLLARY 3.3. *Let $s > 0$, $\epsilon > 0$, $\Gamma > 0$, and $\gamma \in (0,1)$. There exists $N = N(\Gamma, s, \epsilon, \gamma) \in \mathbb{N}^*$ such that, for every $\Psi_0 \in \mathbb{S} \cap (H^2 \cap H^1_0)(I, \mathbb{C})$ with $\Psi_0 \in H^s_{(0)}(I, \mathbb{C})$,*

$$\|\Psi_0\|_{H^s_{(0)}} \leqslant \Gamma \text{ and } |\langle \Psi_0, \phi_1 \rangle| \geqslant \gamma, \tag{3.6}$$

the Cauchy problem (1.1)–(1.3) *with $u(t) = \sigma(\Psi(t)) + v_{\sigma(\psi(t)),N,\epsilon}(\Psi(t))$ has a unique strong solution Ψ. Moreover this solution satisfies* (3.5).

Remark 3. Theorem 1.2 with $\sigma = 0$ is a direct consequence of Corollary 3.3. The feedback law, evoked in Theorem 1.2, is implicitly given by Corollary 3.3.

3.2. Preliminaries.

LEMMA 3.4. *Let $N \in \mathbb{N}^*$, $\epsilon > 0$, and θ satisfy* (3.1). *There exist $C(N) > 0$ and $\tilde{C}(N) > 0$ such that, for all $\xi_1, \xi_2 \in B_{L^2}(0,1)$,*

$$|\sigma(\xi_1) - \sigma(\xi_2)| \leqslant 3N\|\theta'\|_{L^\infty}\|\xi_1 - \xi_2\|_{L^2}, \tag{3.7}$$

$$|\sigma(\xi_1) - \sigma(\xi_2)| \leqslant C(N)\|\theta'\|_{L^\infty}\|\xi_1 - \xi_2\|_{H^{-1}}, \tag{3.8}$$

$$|v_{\sigma(\xi_1),N,\epsilon}(\xi_1) - v_{\sigma(\xi_2),N,\epsilon}(\xi_2)| \leqslant N(1 + 3NC^*\|\theta'\|_{L^\infty})\|\xi_1 - \xi_2\|_{L^2}, \tag{3.9}$$

$$|v_{\sigma(\xi_1),N,\epsilon}(\xi_1) - v_{\sigma(\xi_2),N,\epsilon}(\xi_2)| \leqslant \tilde{C}(N)\|\xi_1 - \xi_2\|_{H^{-1}}. \tag{3.10}$$

Proof of Lemma 3.4. Since N and ϵ are fixed, in order to simplify the notation, we remove them from the subscripts of this proof. We have

$$|\sigma(\xi_1) - \sigma(\xi_2)| \leqslant \|\theta'\|_{L^\infty}|\mathcal{V}_{\sigma(\xi_1)}(\xi_1) - \mathcal{V}_{\sigma(\xi_2)}(\xi_2)|. \tag{3.11}$$

Using

$$\begin{aligned} |\langle \xi_1, \phi_{k,\sigma_1}\rangle|^2 - |\langle \xi_2, \phi_{k,\sigma_2}\rangle|^2 = {} & \langle \xi_1 - \xi_2, \phi_{k,\sigma_1}\rangle \overline{\langle \xi_1, \phi_{k,\sigma_1}\rangle} \\ & + \langle \xi_2, \phi_{k,\sigma_1}\rangle \overline{\langle \xi_1 - \xi_2, \phi_{k,\sigma_1}\rangle} \\ & + \langle \xi_2, \phi_{k,\sigma_1} - \phi_{k,\sigma_2}\rangle \overline{\langle \xi_2, \phi_{k,\sigma_1}\rangle} \\ & + \langle \xi_2, \phi_{k,\sigma_2}\rangle \overline{\langle \xi_2, \phi_{k,\sigma_1} - \phi_{k,\sigma_2}\rangle} \end{aligned} \tag{3.12}$$

and (2.10), we get

$$|\mathcal{V}_{\sigma(\xi_1)}(\xi_1) - \mathcal{V}_{\sigma(\xi_2)}(\xi_2)| \leqslant 2N\|\xi_1 - \xi_2\|_{L^2} + 2NC^*|\sigma(\xi_1) - \sigma(\xi_2)|,$$

$$|\mathcal{V}_{\sigma(\xi_1)}(\xi_1) - \mathcal{V}_{\sigma(\xi_2)}(\xi_2)| \leqslant 2NC_1(N)\|\xi_1 - \xi_2\|_{H^{-1}} + 2NC^*|\sigma(\xi_1) - \sigma(\xi_2)|,$$

where $C_1(N) := \max\{\|\varphi_{k,\sigma}\|_{H^1_0}; k \in \{1, \dots, N\}, \sigma \in [0, \sigma^*]\}$. Using the previous inequalities and (3.1), we get

$$\frac{17}{18}|\sigma(\xi_1) - \sigma(\xi_2)| \leqslant 2N\|\theta'\|_\infty\|\xi_1 - \xi_2\|_{L^2},$$

$$\frac{17}{18}|\sigma_1 - \sigma_2| \leqslant 2NC_1(N)\mathcal{C}\|\theta'\|_\infty\|\xi_1 - \xi_2\|_{H^{-1}},$$

which implies (3.7) and (3.8) with $C(N) = 3N\mathcal{C}C_1(N)$.

Let us write v_j instead of $v_{\sigma(\xi_j)}(\xi_j)$. Using, for the term

$$\langle x\xi_1, \phi_{j,\sigma(\xi_1)}\rangle\overline{\langle \xi_1, \phi_{j,\sigma(\xi_1)}\rangle} - \langle x\xi_2, \phi_{j,\sigma(\xi_2)}\rangle\overline{\langle \xi_2, \phi_{j,\sigma(\xi_2)}\rangle},$$

the same kind of decomposition as in (3.12), together with (2.10), we get

$$|v_1 - v_2| \leqslant N\|\xi_1 - \xi_2\|_{L^2} + NC^*|\sigma(\xi_1) - \sigma(\xi_2)|,$$
$$|v_1 - v_2| \leqslant 2N\mathcal{C}C_1(N)\|\xi_1 - \xi_2\|_{H^{-1}} + 2NC^*|\sigma(\xi_1) - \sigma(\xi_2)|,$$

where $\mathcal{C}$ is defined by (2.21). Thus, using (3.7) and (3.8), we get (3.9) and (3.10) with $\tilde{C}(N) := 2N[\mathcal{C}C_1(N) + C^*C(N)\|\theta'\|_\infty]$. □

PROPOSITION 3.5. *Let $N \in \mathbb{N}^*$, $\epsilon > 0$, and θ verify* (3.1) *and* (3.3). *For every $\Psi_0 \in \mathbb{S}$ the Cauchy problem* (1.1)–(1.3) *with $u(t) = \sigma(\Psi(t)) + v_{\sigma(\psi(t)),N,\epsilon}(\Psi(t))$ has a unique weak solution, i.e., $\Psi \in C^0(\mathbb{R}, \mathbb{S}) \cap C^1((0,+\infty), H^{-2}_{(0)})$. If, moreover, $\Psi \in (H^2 \cap H^1_0)(I, \mathbb{C})$, then Ψ is a strong solution, i.e., $\Psi \in C^0(\mathbb{R}, H^2 \cap H^1_0) \cap C^1((0,+\infty), L^2)$.*

Proof of Proposition 3.5. The strategy is the same as in the proof of Proposition 2.5. Let $T > 0$ be such that

$$NTe^{(N+\|\theta\|_{L^\infty})T} < \frac{1}{2}.$$

Let $\Psi_0 \in \mathbb{S}$. In order to build solutions on $[0,T]$, we apply the Banach fixed-point theorem to the map

$$\begin{array}{rccc} \Theta: & C^0([0,T],\mathbb{S}) & \to & C^0([0,T],\mathbb{S}) \\ & \xi & \mapsto & \Psi, \end{array}$$

where Ψ is the weak solution of (1.1)–(1.3) with $u(t) = \sigma(\xi(t)) + v_{\sigma(\xi(t)),N,\epsilon}(\xi(t))$.

The map Θ is well defined and maps $C^0([0,T],\mathbb{S})$ into itself; moreover, it takes values in $C^0([0,T],\mathbb{S}) \cap C^1((0,T), H^{-2}_{(0)})$ (see Proposition 1.1). Let us prove that Θ is a contraction of $C^0([0,T],\mathbb{S})$. Let $\xi_j \in C^0([0,T],\mathbb{S})$, $v_j := v_{\sigma(\xi_j),N,\epsilon}(\xi_j)$, $\Psi_j := \Theta(\xi_j)$ for $j = 1,2$ and $\Delta := \Psi_1 - \Psi_2$. We have

$$\Delta(t) = i\int_0^t e^{-iA(t-s)}[(\sigma(\xi_1) + v_1)x\Delta(s) + (\sigma(\xi_1) - \sigma(\xi_2) + v_1 - v_2)x\Psi_2(s)]ds.$$

Using (3.7) and (3.9), we get

$$\begin{aligned} \|\Delta(t)\|_{L^2} \leqslant & \int_0^t \Big(\|\theta'\|_{L^\infty} + N\Big)\|\Delta(s)\|_{L^2}ds \\ & + \int_0^t \Big(3N\|\theta'\|_{L^\infty} + N[1 + 3NC^*\|\theta'\|_{L^\infty}]\Big)\|\xi_1 - \xi_2\|_{L^2}ds. \end{aligned}$$

Thus, the Gronwall lemma implies

$$\|\Delta\|_{C^0([0,T],L^2)} \leqslant \|\xi_1 - \xi_2\|_{C^0([0,T],L^2)}[1 + 3(1 + NC^*)\|\theta'\|_{L^\infty}]NTe^{T[N+\|\theta\|_{L^\infty}]}.$$

The choice of T and (3.3) ensures that Θ is a contraction of $C^0([0,T],\mathbb{S})$. Therefore, there exists a fixed point $\Psi \in C^0([0,T],\mathbb{S})$ such that $\Theta(\Psi) = \Psi$. Since Θ takes values in $C^0([0,T],\mathbb{S}) \cap C^1([0,T], H^{-2}_{(0)})$, necessarily Ψ belongs to this space; thus, it is a weak solution of (1.1)–(1.3) on $[0,T]$.

If, moreover, $\Psi_0 \in (H^2 \cap H_0^1)(I, \mathcal{C})$, then the map Θ takes values in $C^0([0,T], H^2 \cap H_0^1) \cap C^1([0,T], L^2)$, and thus Ψ belongs to this space and it is a strong solution.

Since the time T does not depend on Ψ_0, the solution can be continued globally in time. We, therefore, have the existence of global solutions to the closed-loop system. □

Proposition 3.6. *Let $\sigma > 0$, $N \in \mathbb{N}$, $\epsilon > 0$, θ as in* (3.1), $(\Psi_0^n)_{n\in\mathbb{N}}$ *a sequence of* $\mathbb{S}$, *and $\Psi_0^\infty \in L^2$ with $\|\Psi_0^\infty\|_{L^2} \leqslant 1$ such that*

$$\lim_{n\to+\infty} \Psi_0^n = \Psi_0^\infty \textit{ strongly in } H^{-1}(I, \mathbb{C}).$$

Let Ψ^n (resp., Ψ^∞) be the weak solution of (1.1)–(1.3) *with $u(t) = \sigma(\Psi^n(t)) + v_{\sigma(\Psi^n(t)),N,\epsilon}(\Psi^n(t))$ (resp., with $u(t) = \sigma(\Psi^\infty) + v_{\sigma(\Psi^\infty),N,\epsilon}(\Psi^\infty(t))$). Then, for every $\tau > 0$,*

$$\lim_{n\to+\infty} \Psi^n(\tau) = \Psi^\infty(\tau) \textit{ strongly in } H^{-1}(I, \mathbb{C}).$$

Proof of Proposition 3.6. The proof exactly follows that of Proposition 2.6. In order to simplify the notation, we write $v(\Psi)$ instead of $v_{\sigma(\Psi),N,\epsilon}(\Psi)$. We have

$$\begin{aligned}(\Psi^n - \Psi^\infty)(t) = e^{-iAt}(\Psi_0^n - \Psi_0^\infty) &+ i\int_0^t e^{-iA(t-s)}[\sigma(\Psi^n) - \sigma(\Psi^\infty)]x\Psi^n ds \\ &+ i\int_0^t e^{-iA(t-s)}[v(\Psi^n) - v(\Psi^\infty)]x\Psi^n ds \\ &+ i\int_0^t e^{-iA(t-s)}[\sigma(\Psi^\infty) + v(\Psi^\infty)]x(\Psi^n - \Psi^\infty) ds.\end{aligned}$$

Using (3.8), (3.10), and $\|x\Psi\|_{H^{-1}} \leqslant \|x\Psi\|_{L^2} \leqslant 1$, we get

$$\begin{aligned}&\|(\Psi^n - \Psi^\infty)(t)\|_{H^{-1}} \leqslant \|\Psi_0^n - \Psi_0^\infty\|_{H^{-1}} \\ &+ \int_0^t \Big(C(N)\|\theta'\|_{L^\infty} + \tilde{C}(N) + \mathcal{C}(\|\theta\|_{L^\infty} + N)\Big)\|\Psi^n - \Psi^\infty\|_{H^{-1}} ds,\end{aligned}$$

where $\mathcal{C}$ is given by (2.21). The Gronwall lemma concludes the proof. □

3.3. Proofs of Theorem 3.2 and Corollary 3.3.

Proof of Theorem 3.2. For $\varphi \in B_{L^2}(0,2)$, we define

$$\mathcal{V}_{N,\epsilon}(\varphi) := \mathcal{V}_{\sigma(\varphi),N,\epsilon}(\varphi),$$

where $\mathcal{V}_{\sigma,N,\epsilon}$ is defined by (2.23). Since N and ϵ are fixed, in order to simplify the notation, we omit them in the subscripts of this proof and write $v(\Psi)$ instead of $v_{\sigma(\Psi),N,\epsilon}(\Psi)$.

Let $\Psi_0 \in \mathbb{S} \cap (H^2 \cap H_0^1)(I, \mathbb{C})$ and let Ψ be the strong solution of (1.1)–(1.3) with $u(t) = \sigma(\Psi(t)) + v_{\sigma(\psi(t)),N,\epsilon}(\Psi(t))$ given by Proposition 3.5. Since $\Psi \in C^1(\mathbb{R}, L^2)$ and $\sigma \in C^\infty(B_{L^2}(0,2))$, the map $t \mapsto \mathcal{V}(\psi(t))$ is C^1. We have

$$\frac{d}{dt}\mathcal{V}(\Psi) = -2v(\Psi)^2 - \frac{d}{dt}\Big[\sigma(\Psi)\Big]\Re\left(\sum_{k=1}^N a_k \left\langle \Psi, \frac{d\phi_{k,\sigma}}{d\sigma}\Big|_{\sigma(\Psi)}\right\rangle \overline{\langle \Psi, \phi_{k,\sigma(\Psi)}\rangle}\right),$$

where $a_1 := 1$ and $a_k := 1 - \epsilon$ for $k = 2, \ldots, N$. Moreover,

$$\frac{d}{dt}\Big[\sigma(\Psi)\Big] = \theta'(\mathcal{V}(\psi))\frac{d}{dt}\mathcal{V}(\Psi),$$

and thus

(3.13)
$$\left[1+2\theta'(\mathcal{V}(\psi))\Re\left(\sum_{k=1}^{N} a_k\left\langle \Psi, \frac{d\phi_{k,\sigma}}{d\sigma}\Big|_{\sigma(\Psi)}\right\rangle\overline{\langle \Psi, \phi_{k,\sigma(\Psi)}\rangle}\right)\right]\frac{d}{dt}\mathcal{V}(\Psi) = -2v(\Psi)^2.$$

Using (2.8) and (3.1), we get

$$1+2\theta'(\mathcal{V}(\psi))\Re\left(\sum_{k=1}^{N} a_k\left\langle \Psi, \frac{d\phi_{k,\sigma}}{d\sigma}\Big|_{\sigma(\Psi)}\right\rangle\overline{\langle \Psi, \phi_{k,\sigma(\Psi)}\rangle}\right) \geqslant 1-2\|\theta'\|_{L^\infty}NC^* > 0;$$

thus, $t\mapsto \mathcal{V}(\Psi(t))$ is a nonincreasing function. There exists $\alpha \in [0, \mathcal{V}(\Psi_0)]$ such that

$$\lim_{t\to+\infty} \mathcal{V}(\Psi(t)) = \alpha.$$

Using (2.10), (3.2), and (3.4), we get

$$\begin{aligned}
|\langle \Psi_0, \phi_{1,\sigma(\Psi_0)}\rangle| &\geqslant |\langle \Psi_0, \phi_1\rangle| - |\langle \Psi_0, \phi_1 - \phi_{1,\sigma(\Psi_0)}\rangle| \\
&\geqslant \gamma - C^*\|\theta\|_\infty \\
&\geqslant \tilde{\gamma} := \frac{\gamma}{2},
\end{aligned}$$

$$\begin{aligned}
\sum_{k=N+1}^{\infty} |\langle \Psi_0, \phi_{k,\sigma(\Psi_0)}\rangle|^2 &\leqslant 2\sum_{k=N+1}^{\infty}\left(|\langle \Psi_0, \phi_k\rangle|^2 + |\langle \Psi_0, \phi_{k,\sigma(\Psi_0)} - \phi_k\rangle|^2\right) \\
&\leqslant \frac{\epsilon\gamma^2}{16(1-\epsilon/2)} + 2(C^*\|\theta\|_{L^\infty})^2 \sum_{k=N+1}^{\infty}\frac{1}{k^2} \\
&\leqslant \frac{\epsilon\gamma^2}{16(1-\epsilon/2)} + \frac{2(C^*\|\theta\|_{L^\infty})^2}{N} \\
&\leqslant \frac{\tilde{\epsilon}\tilde{\gamma}^2}{(1-\tilde{\epsilon})},
\end{aligned}$$

where $\tilde{\epsilon} := \epsilon/2$. Thus, as in the proof of Theorem 2.1, $\mathcal{V}(\Psi_0) < \tilde{\epsilon}$, so $\alpha \in (0, \tilde{\epsilon})$.

Let $(t_n)_{n\in\mathbb{N}}$ be an increasing sequence of positive real numbers such that $t_n \to +\infty$ when $n \to +\infty$. Since $\|\Psi(t_n)\|_{L^2} = 1$ for every $n \in \mathbb{N}$, there exists $\Psi_\infty \in L^2(I,\mathbb{C})$ such that, up to an extraction,

$$\Psi(t_n) \to \Psi_\infty \text{ weakly in } L^2(I,\mathbb{C}) \text{ and strongly in } H^{-1}(I,\mathbb{C}).$$

Let ξ be the weak solution of

$$\begin{cases}
i\frac{\partial\xi}{\partial t} = A_\sigma\xi - v_{\sigma(\xi),N,\epsilon}(\xi(t))x\xi, \\
\xi(t,\pm 1/2) = 0, \\
\xi(0) = \Psi_\infty.
\end{cases}$$

Thanks to Proposition 3.6, for every $\tau > 0$, $\Psi(t_n+\tau) \to \xi(\tau)$ strongly in $H^{-1}(I,\mathbb{C})$ when $n\to+\infty$, and thus $\sigma(\Psi(t_n+\tau)) \to \sigma(\xi(\tau))$ when $n \to +\infty$ (see Lemma 3.4). Therefore, $\mathcal{V}(\Psi(t_n+\tau)) \to \mathcal{V}(\xi(\tau))$ when $n\to+\infty$, so $\mathcal{V}(\xi)\equiv\alpha$. Thus, $\sigma(\xi)\equiv\overline{\sigma} := \theta(\alpha)$ and we have, for every $t\in\mathbb{R}_+$,

$$\mathcal{V}(\xi(t)) = 1 - |\langle \xi(t), \phi_{1,\overline{\sigma}}\rangle|^2 - (1-\epsilon)\sum_{k=2}^{N}|\langle \xi(t), \phi_{k,\overline{\sigma}}\rangle|^2.$$

Since $\xi \in C^1(\mathbb{R}_+, H^{-2}_{(0)})$, the previous equality implies

$$\frac{d\mathcal{V}(\xi)}{dt} = -2v(\xi)^2.$$

Since $\mathcal{V}(\xi) \equiv \alpha$, then $v(\xi) \equiv 0$.

First case: $\alpha = 0$. Then $\mathcal{V}(\Psi(t)) \to 0$ when $t \to +\infty$ and $\overline{\sigma} = 0$. Moreover, for every $t \in (0, \infty)$,

$$\begin{aligned}\mathcal{V}(\Psi(t)) &\geqslant 1 - |\langle \Psi, \phi_{1,\sigma(\Psi)}\rangle|^2 - (1-\epsilon)\sum_{k=2}^{\infty} |\langle \Psi, \phi_{k,\sigma(\Psi)}\rangle|^2 \\ &\geqslant \epsilon(1 - |\langle \Psi, \phi_{1,\sigma(\Psi)}\rangle|^2).\end{aligned}$$

Thus,

$$|\langle \Psi(t_n), \phi_{1,\sigma(\Psi(t_n))}\rangle| \to 1,$$

which leads to

$$|\langle \Psi(t_n), \phi_1\rangle| \to 1$$

because $\sigma(\Psi(t_n)) \to 0$.

Second case: $\alpha \neq 0$. Then $\overline{\sigma} = \theta(\alpha) > 0$. Exactly as in the first analysis, done in the proof of Theorem 2.1, we get

$$\Psi_\infty = \beta\phi_{1,\overline{\sigma}},$$

where $\beta \in \mathbb{C}$ and $|\beta|^2 > 1 - \tilde{\epsilon}$. Thus

$$\lim_{n\to+\infty} |\langle \Psi(t_n), \phi_1\rangle| = |\langle \Psi_\infty, \phi_1\rangle| \geqslant |\beta| - |\langle \Psi_\infty, \phi_{1,\overline{\sigma}} - \phi_1\rangle| \geqslant \sqrt{1-\epsilon/2} - C^*\overline{\sigma},$$

where we used (2.7) in the last inequality. Finally, thanks to $0 < \overline{\sigma} \leqslant \|\theta\|_\infty$ and (3.2), we get (3.5). □

Proof of Corollary 3.3. It can be done in a very similar way to the proof of Corollary 2.2. □

4. Numerical simulations. In this section, we check out the performance of the techniques on some numerical simulations. We consider, as a test case, the stabilization of the initial state $\Psi_0 = \frac{1}{\sqrt{2}}(\phi_{1,\sigma} + \phi_{3,\sigma})$ around the ground state $\phi_{1,\sigma}$. Therefore, the cut-off dimension N is 3. Note that such a test case is a particularly hard one in a near-degenerate situation. Indeed, considering the feedback law (1.9) for $\sigma = 0$, one can easily see that for parity reasons $v(\Psi(t)) \equiv 0$.

In a first simulation, we consider the nondegenerate case of $\sigma \neq 0$. As mentioned above, the constant σ needs to be small. In fact, one should choose σ, such that the perturbation σx is small compared to the operator $-\frac{1}{2}\frac{\partial^2}{\partial x^2}$. We choose it here to be $\sigma = 2e+01$. Figure 4.1 illustrates the simulation of the closed-loop system when $u = \sigma + v_\epsilon$ with $\varsigma = 1e+03$ and $\epsilon = 5e-02$. The simulations have been done applying a third order split-operator method (see, e.g., [13]), where instead of computing $\exp(-i\ dt\ (A_\sigma - v_\epsilon x))$ at each time step, we compute

$$\exp(-i\ dt\ A_\sigma/2)\exp(i\ dt\ v_\epsilon x)\exp(-i\ dt\ A_\sigma/2).$$

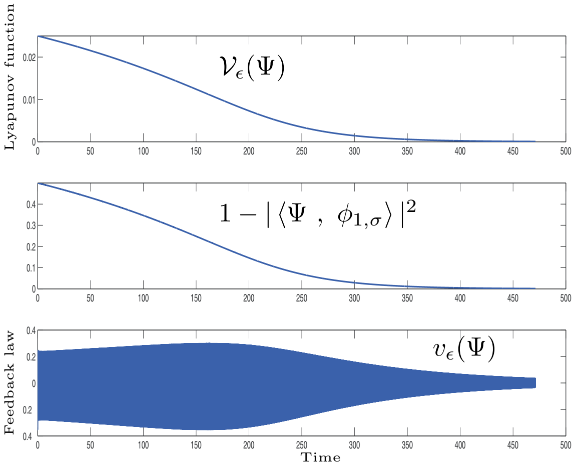


FIG. 4.1. *The practical stabilization of* $\mathcal{C}_{1,\sigma}$*, where* $\Psi_0 = \frac{1}{\sqrt{2}}(\phi_{1,\sigma} + \phi_{3,\sigma})$ *and therefore the cut-off dimension is* 3*; as can be seen, the closed-loop system reaches the* .05*-neighborhood of* $\phi_{1,\sigma}$ *in a time* $T = 150\pi$ *corresponding to about* 200 *periods of the longest natural period corresponding to the ground to the first excited state.*

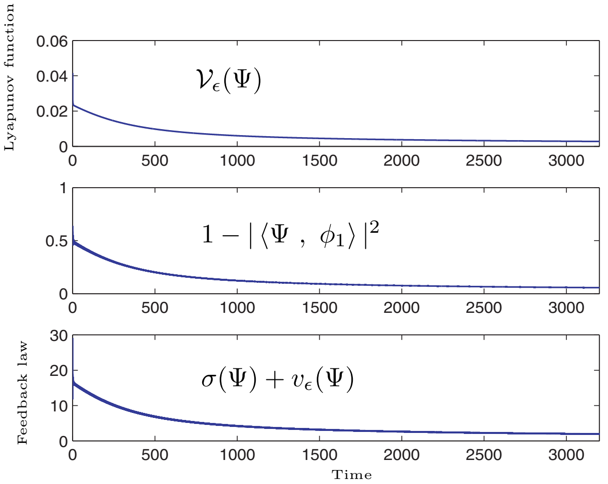


FIG. 4.2. *The practical stabilization of* $\mathcal{C}_1$*, where* $\Psi_0 = \frac{1}{\sqrt{2}}(\phi_1 + \phi_3)$ *and therefore the cut-off dimension is* 3*; as can be seen, the closed-loop system reaches the* .05*-neighborhood of* ϕ_1 *in a time* $T = 1000\pi$ *corresponding to about* 1300 *periods of the longest natural period corresponding to the ground to the first excited state.*

Moreover, we consider a Galerkin discretization over the first 20 modes of the system (it turns out, by considering higher modal approximations, that 20 modes are completely sufficient to get a reliable result).

Now, let us consider the degenerate case of $\sigma = 0$. As mentioned above, such a case is not treatable with the explicit feedback design of (1.9). However, the simulations of Figure 4.2 show that the implicit Lyapunov design provided in subsection 1.3 removes the degeneracy problem and ensures the practical stabilization of the initial state $\frac{1}{\sqrt{2}}(\phi_1 + \phi_3)$ around the ground state ϕ_1.

We consider the function $\theta(r) = \eta r$ with $\eta = 7e{+}02$. Furthermore, in the feedback design v_ϵ, we consider $\varsigma = 1e+03$ and $\epsilon = 5e-02$. The numerical scheme is similar to the simulations of Figure 4.1. In order to calculate the implicit part of the feedback design $\sigma(\Psi)$, we apply a fixed-point algorithm.

5. Appendix.

This appendix is devoted to the proofs of Propositions 1.1 and 3.1.

5.1. Proof of Proposition 1.1.

Let $\Psi_0 \in \mathbb{S}$, $T_1 > 0$ and $u \in C^0([0,T_1],\mathbb{R})$. Let $T \in (0,T_1)$ be such that

$$\|u\|_{L^1(0,T)} < 1. \tag{5.1}$$

We prove the existence of $\Psi \in C^0([0,T],L^2(I,\mathbb{C}))$ such that (1.4) holds by applying the Banach fixed-point theorem to the map

$$\begin{array}{rccc} \Theta: & C^0([0,T],L^2) & \rightarrow & C^0([0,T],L^2) \\ & \xi & \mapsto & \Psi, \end{array}$$

where Ψ is the weak solution of

$$\begin{cases} i\frac{\partial\Psi}{\partial t} = A\Psi - u(t)x\xi, \\ \Psi(0,x) = \Psi_0(x), \\ \Psi(t,\pm 1/2) = 0, \end{cases}$$

i.e., $\Psi \in C^0([0,T],L^2)$ and satisfies, for every $t \in [0,T]$,

$$\Psi(t) = e^{-iAt}\Psi_0 + i\int_0^t e^{-iA(t-s)}u(s)x\xi(s)ds \text{ in } L^2(I,\mathbb{C}).$$

Notice that Θ takes values in $C^1([0,T],H^{-2}_{(0)}(I,\mathbb{C}))$.

For $\xi_1,\xi_2 \in C^0([0,T],L^2(I,\mathbb{C}))$, $\Psi_1 := \Theta(\xi_1)$, $\Psi_2 := \Theta(\xi_2)$ we have

$$(\Psi_1 - \Psi_2)(t) = i\int_0^t e^{-iA(t-s)}u(s)x(\xi_1 - \xi_2)(s)ds,$$

and thus

$$\|(\Psi_1 - \Psi_2)(t)\|_{L^2} \leqslant \int_0^t |u(s)|ds\|\xi_1 - \xi_2\|_{C^0([0,T],L^2)}.$$

The assumption (5.1) guarantees that Θ is a contraction of $C^0([0,T],L^2)$, and thus Θ has a fixed point $\Psi \in C^0([0,T],L^2)$. Since Θ takes values in $C^1([0,T],H^{-2}_{(0)})$, then Ψ belongs to this space. Moreover, this function satisfies (1.4).

Finally, we have built weak solutions on $[0,T]$ for every Ψ_0, and the time T does not depend on Ψ_0. Thus, this gives solutions on $[0,T_1]$.

Let us prove that this solution is continuous with respect to the the initial condition Ψ_0 for the $L^2(I,\mathbb{C})$-topology. Let $\Psi_0, \Phi_0 \in \mathbb{S}$ and let Ψ, Φ be the associated weak solutions. We have

$$\|(\Psi - \Phi)(t)\|_{L^2} \leqslant \|\Psi_0 - \Phi_0\|_{L^2} + \int_0^t |u(s)|\|(\Psi - \Phi)(s)\|_{L^2}ds,$$

and thus the Gronwall lemma gives

$$\|(\Psi - \Phi)(t)\|_{L^2} \leqslant \|\Psi_0 - \Phi_0\|_{L^2} e^{\|u\|_{L^1(0,T_1)}}.$$

This gives the continuity of the weak solutions with respect to the initial conditions.

Now, let us assume that $\Psi_0 \in H^2 \cap H^1_0(I, \mathbb{C})$. Take C to be a positive constant such that for every $\varphi \in H^2 \cap H^1_0(I, \mathbb{C})$, $\|x\varphi\|_{H^2 \cap H^1_0} \leqslant C\|\varphi\|_{H^2 \cap H^1_0}$. We consider, then, $T > 0$ such that $C\|u\|_{L^1(0,T)} < 1$. By applying the fixed-point theorem on

$$\Theta_2 : C^0([0,T], H^2 \cap H^1_0) \to C^0([0,T], H^2 \cap H^1_0)$$

defined by the same expression as Θ, and using the uniqueness of the fixed point of Θ, we get that the weak solution is a strong solution. The continuity with respect to the initial condition of the strong solution can also be proved by applying the same arguments as in above.

Finally, let us justify that the weak solutions take values in $\mathbb{S}$. For $\Psi_0 \in H^2 \cap H^1_0$, the solution belongs to $C^1([0,T], L^2) \cap C^0([0,T], H^2 \cap H^1_0)$ and thus the following computations are justified:

$$\frac{d}{dt}\|\Psi(t)\|^2_{L^2} = 2\Re\left\langle \frac{\partial \Psi}{\partial t}, \Psi \right\rangle = 0.$$

Thus $\Psi(t) \in \mathbb{S}$ for every $t \in [0,T]$.

For $\Psi_0 \in \mathbb{S}$, we get the same conclusion thanks to a density argument and the continuity for the $C^0([0,T], L^2)$-topology of the weak solutions with respect to the initial condition. □

5.2. Proof of Proposition 3.1. Let $\Psi \in B_{L^2}(0,2)$. We prove the existence of $\sigma(\Psi)$ by applying the Banach fixed-point theorem to the map

$$\begin{array}{rccc} \Pi : & [0, \|\theta\|_{L^\infty}] & \to & [0, \|\theta\|_{L^\infty}] \\ & \sigma & \mapsto & \theta(\mathcal{V}_{\sigma,N,\epsilon}(\Psi)). \end{array}$$

For $\sigma_1, \sigma_2 \in [0, \|\theta\|_{L^\infty}]$, we have

$$|\Pi(\sigma_1) - \Pi(\sigma_2)| \leqslant \|\theta'\|_{L^\infty} |\mathcal{V}_{\sigma_1,N,\epsilon}(\Psi) - \mathcal{V}_{\sigma_2,N,\epsilon}(\Psi)|.$$

Using the inequality

$$\begin{aligned} \left| |\langle \Psi, \phi_{j,\sigma_1}\rangle|^2 - |\langle \Psi, \phi_{j,\sigma_2}\rangle|^2 \right| &\leqslant \left| \langle \Psi, \phi_{j,\sigma_1} - \phi_{j,\sigma_2}\rangle \overline{\langle \Psi, \phi_{j,\sigma_1}\rangle} \right| \\ &\quad + \left| \langle \Psi, \phi_{j,\sigma_2}\rangle \overline{\langle \Psi, \phi_{j,\sigma_1} - \phi_{j,\sigma_2}\rangle} \right| \\ &\leqslant 8\|\phi_{j,\sigma_1} - \phi_{j,\sigma_2}\|_{L^2}, \end{aligned}$$

together with (2.10), we get

$$|\Pi(\sigma_1) - \Pi(\sigma_2)| \leqslant 8NC^* \|\theta'\|_{L^\infty} |\sigma_1 - \sigma_2|.$$

Thus, the assumption (3.1) ensures that Π is a contraction of $[0, \|\theta\|_{L^\infty}]$. Therefore, Π has a unique fixed point $\sigma(\Psi)$.

Now, let us prove that σ is C^∞. The map

$$\begin{array}{rccc} F : & [0, \|\theta\|_{L^\infty}] \times B_{L^2}(0,2) & \to & \mathbb{R} \\ & (\sigma, \Psi) & \mapsto & \sigma - \theta(\mathcal{V}_{\sigma,N,\epsilon}(\Psi)) \end{array}$$

is regular with respect to σ and Ψ, $F(\sigma(\Psi),\Psi)=0$ for every $\Psi\in B_{L^2}(0,2)$, and

$$\frac{\partial F}{\partial\sigma}(\sigma(\Psi),\Psi)=1-2\theta'(\mathcal{V}_{\sigma(\Psi),N,\epsilon}(\Psi))\frac{\partial}{\partial\sigma}\Big[\mathcal{V}_{\sigma,N,\epsilon}(\Psi)\Big]_{\sigma(\psi)}\geqslant\frac{1}{2}. \tag{5.2}$$

Indeed, for $\sigma_0\in[0,\|\theta\|_{L^\infty}]$ and $\Psi\in B_{L^2}(0,2)$, we have

$$\frac{\partial}{\partial\sigma}\Big[\mathcal{V}_{\sigma,N,\epsilon}(\Psi)\Big]_{\sigma_0}=-2\sum_{k=1}^{N}a_k\Re\left(\left\langle\Psi,\left.\frac{d\phi_{k,\sigma}}{d\sigma}\right|_{\sigma_0}\right\rangle\overline{\langle\Psi,\phi_{k,\sigma_0}\rangle}\right),$$

where $a_1:=1$ and $a_k:=1-\epsilon$ for $k=2,\dots,N$. Thus, using (2.8), we get

$$\left|\frac{\partial}{\partial\sigma}\Big[\mathcal{V}_{\sigma,N,\epsilon}(\Psi)\Big]_{\sigma_0}\right|\leqslant 8NC^*.$$

We get the inequality in (5.2) thanks to the previous inequality and (3.1).

For every $\Psi\in B_{L^2}(0,2)$, the implicit function theorem provides the existence of a local C^∞ parameterization $\tilde\sigma(\xi)$ for the solutions of $F(\sigma(\xi),\xi)=0$ in a neighborhood of Ψ. The uniqueness of the fixed point $\sigma(\xi)$ justifies that σ and $\tilde\sigma$ coincide, and thus σ is C^∞. □

Acknowledgments. The authors thank J.-M. Coron, R. van Handel, O. Kavian, and P. Rouchon for helpful discussions.

REFERENCES

[1] F. Albertini and D. D'Alessandro, *Notions of controllability for bilinear multilevel quantum systems*, IEEE Trans. Automat. Control, 48 (2003), pp. 1399–1403.

[2] C. Altafini, *Controllability of quantum mechanical systems by root space decomposition of su(n)*, J. Math. Phys., 43 (2002), pp. 2051–2062.

[3] L. Baudouin and J. Salomon, *Constructive solutions of a bilinear control problem*, C. R. Math. Acad. Sci. Paris, 342 (2006), pp. 119–124.

[4] L. Baudouin, O. Kavian, and J.-P. Puel, *Regularity for a Schrödinger equation with singular potentials and application to bilinear optimal control*, J. Differential Equations, 216 (2005), pp. 188–222.

[5] K. Beauchard, *Local controllability of a 1-D Schrödinger equation*, J. Math. Pures Appl., 84 (2005), pp. 851–956.

[6] K. Beauchard and J. M. Coron, *Controllability of a quantum particle in a moving potential well*, J. Funct. Anal., 232 (2006), pp. 328–389.

[7] K. Beauchard, J.-M. Coron, M. Mirrahimi, and P. Rouchon, *Implicit Lyapunov control of finite dimensional Schrödinger equations*, System Control Lett., 56 (2007), pp. 388–395.

[8] Y. Chen, P. Gross, V. Ramakrishna, H. Rabitz, and K. Mease, *Competitive tracking of molecular objectives described by quantum mechanics*, J. Chem. Phys., 102 (1995), pp. 8001–8010.

[9] J. M. Coron, *Global stabilization for controllable systems without drift*, Math. Control Signals Systems, 5 (1992), pp. 295–312.

[10] J. M. Coron, *Control and Nonlinearity*, Math. Surveys Monogr. 136, AMS, Providence, RI, 2007.

[11] J.-M. Coron and B. d'Andréa-Novel, *Stabilization of a rotating body-beam without damping*, IEEE Trans. Automat. Control, 43 (1998), pp. 608–618.

[12] A. Elgart and J. E. Avron, *Adiabatic theorem without a gap condition*, Commun. Math. Phys., 203 (1999), pp. 445–463.

[13] M. D. Feit, J. A. Fleck, Jr., and A. Steiger, *Solution of the Schrödinger equation by a spectral method*, J. Comput. Phys., 47 (1982), pp. 412–433.

[14] S. Haroche, *Contrôle de la décohérence: Théorie et expériences*, Notes de cours, Collège de France, 2004; http://www.lkb.ens.fr/recherche/qedcav/college/college.html.

[15] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.

[16] W. Krabs, *On Moment Theory and Controllability of One Dimensional Vibrating Systems and Heating Processes*, Springer-Verlag, Berlin, 1992.

[17] B. Li, G. Turinici, V. Ramakrishna, and H. Rabitz, *Optimal dynamic discrimination of similar molecules through quantum learning control*, J. Phys. Chem. B, 106 (2002), pp. 8125–8131.

[18] M. Mirrahimi, *Lyapunov control of a quantum particle in a decaying potential*, Ann. Inst. H. Poincaré Anal. Non Linéaire, to appear.

[19] M. Mirrahimi and R. Van Handel, *Stabilizing feedback controls for quantum systems*, SIAM J. Control Optim., 46 (2007), pp. 445–467.

[20] M. Mirrahimi, P. Rouchon, and G. Turinici, *Lyapunov control of bilinear Schrödinger equations*, Automatica, 41 (2005), pp. 1987–1994.

[21] M. Mirrahimi, G. Turinici, and P. Rouchon, *Reference trajectory tracking for locally designed coherent quantum controls*, J. Phys. A, 109 (2005), pp. 2631–2637.

[22] V. Ramakrishna, M. Salapaka, M. Dahleh, and H. Rabitz, *Controllability of molecular systems*, Phys. Rev. A, 51 (1995), pp. 960–966.

[23] P. Rouchon, *Control of a Quantum Particle in a Moving Box*, Technical Report A/325, École des Mines de Paris, Centre Automatique et Systèmes, 2002.

[24] S. Shi, A. Woody, and H. Rabitz, *Optimal control of selective vibrational excitation in harmonic linear chain molecules*, J. Chem. Phys., 88 (1988), pp. 6870–6883.

[25] M. Sugawara, *General formulation of locally designed coherent control theory for quantum systems*, J. Chem. Phys., 118 (2003), pp. 6784–6800.

[26] H. J. Sussmann and V. Jurdjevic, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95–116.

[27] G. Turinici, *Controllable quantities for bilinear quantum systems*, in Proceedings of the 39th IEEE Conference on Decision and Control, 2000, pp. 1364–1369.

[28] G. Turinici and H. Rabitz, *Wavefunction controllability in quantum systems*, J. Phys. A, 36 (2003), pp. 2565–2576.

[29] R. Van Handel, J. K. Stockton, and H. Mabuchi, *Modeling and feedback control design for quantum state preparation*, J. Opt. B Quantum Semiclass. Opt., 7 (2005), pp. S179–S197.

SIAM J. CONTROL OPTIM.
Vol. 48, No. 2, pp. 1206–1228

© 2009 Society for Industrial and Applied Mathematics

ANALYSIS AND DESIGN OF UNCONSTRAINED NONLINEAR MPC SCHEMES FOR FINITE AND INFINITE DIMENSIONAL SYSTEMS[*]

LARS GRÜNE[†]

Abstract. We present a technique for computing stability and performance bounds for unconstrained nonlinear model predictive control (MPC) schemes. The technique relies on controllability properties of the system under consideration, and the computation can be formulated as an optimization problem whose complexity is independent of the state space dimension. Based on the insight obtained from the numerical solution of this problem, we derive design guidelines for nonlinear MPC schemes which guarantee stability of the closed loop for small optimization horizons. These guidelines are illustrated by a finite and an infinite dimensional example.

Key words. model predictive control, suboptimality, stability, controllability, linear programming, controller design, infinite dimensional system

AMS subject classifications. 49N35, 93D15, 93B05

DOI. 10.1137/070707853

1. Introduction.

Model predictive control (MPC, often also termed receding horizon control) is a well-established method for the optimal control of linear and nonlinear systems [1, 2, 19]. The stability and suboptimality analysis of MPC schemes has been a topic of active research during the last decades. In the MPC literature, in order to prove stability and suboptimality of the resulting closed loop, often stabilizing terminal constraints or terminal costs are used (see, e.g., [14], [3], [11], [17], or the survey paper [19]). In these schemes, the terminal costs need to satisfy Lyapunov function-type conditions, as, e.g., condition [3, Equation (9)] in the case of terminal constraints and [17, Equations (12) and (13)] in the case without terminal constraints. In contrast to these approaches, here we consider the simplest class of MPC schemes for nonlinear systems, namely, those without terminal constraints and cost, which we call "unconstrained" MPC schemes. These schemes are attractive for their numerical simplicity, do not require the computation of Lyapunov function like terminal costs or the introduction of stabilizing state space constraints—which are particularly inconvenient when treating infinite dimensional systems—and are easily generalized to time-varying, tracking-type problems and to the case where more complicated sets than equilibria are to be stabilized. Furthermore, they appear to be the class of schemes most commonly used in industrial practice, cf. [22, section 7]. Essentially, these unconstrained MPC schemes can be interpreted as a simple truncation of the infinite optimization horizon to a finite horizon N.

For such unconstrained schemes without terminal cost, Jadbabaie and Hauser [13] and Grimm et al. [4] show under different types of controllability and detectability conditions for nonlinear systems that stability of the closed loop can be expected if the optimization horizon N is sufficiently large; however, no explicit bounds for N are given. The paper [8] (see also [7] and [6] for variants of this approach) uses controllability conditions and techniques from relaxed dynamic programming [16, 23] in

[*] Received by the editors November 9, 2007; accepted for publication (in revised form) December 3, 2008; published electronically March 25, 2009. This research was supported by the Deutsche Forschungsgemeinschaft (DFG) grant Gr1569/12-1.

http://www.siam.org/journals/sicon/48-2/70785.html

[†] Mathematisches Institut, Universität Bayreuth, 95440 Bayreuth, Germany (lars.gruene@uni-bayreuth.de).

order to compute explicit estimates for the degree of suboptimality, which, in particular, lead to bounds on the stabilizing optimization horizon N which are, however, in general not optimal. Such optimal estimates for the stabilizing horizon N have been obtained in [24, 21] using the explicit knowledge of the finite horizon optimal value functions, which could be computed numerically in the (linear) examples considered in these papers.

Unfortunately, for large scale or infinite dimensional and also for moderately sized nonlinear systems in general neither an analytical expression nor a sufficiently accurate numerical approximation of optimal value functions is available. Furthermore, an analysis based on such numerical approximations typically does not provide analytic insight into the dependence between the stability properties and the system structure. For these reasons, in this paper we base our analysis on (open loop) controllability properties, which can often be estimated or characterized in sufficient detail by analyzing the system structure. More precisely, for our analysis we use $\mathcal{KL}$ bounds of the chosen running cost along (not necessarily optimal) trajectories. Such bounds induce upper bounds on the optimal value functions and the main feature we exploit is the fact that the controllability properties do not only impose bounds on the optimal value function at the initial value but—via Bellman's optimality principle—also along "tails" of optimal trajectories. Extending preliminary results in this direction from [5], we show that the resulting stability and suboptimality condition can be expressed as an optimization problem whose complexity is independent of the dimension of the state space of the system and which is actually an easily solvable linear program if the $\mathcal{KL}$ function involved in the controllability assumption is linear in its first argument. As in [8], this procedure gives a bound on the degree of suboptimality of the MPC feedback which, in particular, allows one to determine a bound on the minimal stabilizing horizon N, but in contrast to [8] the bound derived here turns out to be optimal with respect to the class of systems satisfying the assumed controllability property.

Since the resulting optimization problem is small and, thus, easy to solve, we can perform a comprehensive numerical analysis of many different controllability situations, which we use in order to derive design guidelines for the formulation of stable MPC schemes with small optimization horizon N. A distinctive feature of our approach is that our analysis applies to finite and infinite dimensional systems alike and we demonstrate the effectiveness of our approach in an infinite dimensional setting by an example of a sampled data system governed by a parabolic PDE.

The paper is organized as follows: in section 2 we describe the setup and the relaxed dynamic programming inequality our approach is based upon. In section 3 we describe the controllability condition we are going to use and its consequences to the optimal value functions and trajectories. In section 4 we use these results in order to obtain a condition for suboptimality and show how this condition can be formulated as an optimization problem. Section 5 shows how our condition can be used for the closed loop stability analysis. In section 6 we perform a case study in which we analyze the impact of different controllability bounds and MPC parameters on the minimal stabilizing horizon N. Based on the numerical findings from this analysis, in section 7 we formulate our design guidelines for MPC schemes and illustrate them by two examples. We finish the paper by giving conclusions and outlook in section 8 and the formulation and proof of a technical lemma in the Appendix.

2. Setup and preliminary results.

We consider a nonlinear discrete time system given by

$$x(n+1) = f(x(n), u(n)), \quad x(0) = x_0 \tag{2.1}$$

with $x(n) \in X$ and $u(n) \in U$ for $n \in \mathbb{N}_0$. Here we denote the space of control sequences $u : \mathbb{N}_0 \to U$ by $\mathcal{U}$ and the solution trajectory for some $u \in \mathcal{U}$ by $x_u(n)$. The state space X is an arbitrary metric space; i.e., it can range from a finite set to an infinite dimensional space.

A typical class of systems we consider are sampled-data systems governed by a controlled—finite or infinite dimensional—differential equation $\dot{x}(t) = g(x(t), \tilde{u}(t))$ with solution $\varphi(t, x_0, \tilde{u})$ for initial value x_0. These are obtained by fixing a sampling period $T > 0$ and setting

$$f(x,u) := \varphi(T, x, \tilde{u}) \quad \text{with} \quad \tilde{u}(t) \equiv u. \tag{2.2}$$

Then, for any discrete time control function $u \in \mathcal{U}$ the solutions x_u of (2.1), (2.2) satisfy $x_u(n) = \varphi(nT, x_0, \tilde{u})$ for the piecewise constant continuous time control function $\tilde{u} : \mathbb{R} \to U$ with $\tilde{u}|_{[nT,(n+1)T)} \equiv u(n)$. Note that with this construction the discrete time n corresponds to the continuous time $t = nT$.

Our goal is to find a feedback control law minimizing the infinite horizon cost

$$J_\infty(x_0, u) = \sum_{n=0}^{\infty} l(x_u(n), u(n)), \tag{2.3}$$

with running cost $l : X \times U \to \mathbb{R}_0^+$. We denote the optimal value function for this problem by

$$V_\infty(x_0) = \inf_{u \in \mathcal{U}} J_\infty(x_0, u).$$

Here we use the term feedback control in the following general sense.

Definition 2.1. *For $m \geq 1$, an m-*step feedback law *is a map $\mu : X \times \{0, \dots, m-1\} \to U$ which is applied according to the rule*

$$x_\mu(n+1) = f(x_\mu(n), \mu(x_\mu([n]_m), n - [n]_m)), \quad x_\mu(0) = x_0, \tag{2.4}$$

where $[n]_m$ denotes the largest product km, $k \in \mathbb{Z}$, with $km \leq n$.

In other words, the feedback is evaluated at the times $0, m, 2m \dots$ and generates a sequence of m control values which is applied in the m steps until the next evaluation. Note that for $m = 1$ we obtain the usual static state feedback concept in discrete time.

If the optimal value function V_∞ is known, it is easy to prove using Bellman's optimality principle that the optimal feedback law μ is given by

$$\mu(x_0, \cdot) := \operatorname*{argmin}_{u \in U^m} \left\{ V_\infty(x_u(m)) + \sum_{n=0}^{m-1} l(x_u(n), u(n)) \right\}. \tag{2.5}$$

Remark 2.2. We assume throughout this paper that in all relevant expressions the minimum with respect to $u \in U^m$ is attained. Although it is possible to give modified statements using approximate minimizers, we decided to make this assumption in order to simplify and streamline the presentation.

Since infinite horizon optimal control problems are in general computationally infeasible, we use a receding horizon approach in order to compute an approximately optimal controller. To this end we consider the finite horizon functional

$$J_N(x_0, u) = \sum_{n=0}^{N-1} l(x_u(n), u(n)) \tag{2.6}$$

for $N \in \mathbb{N}_0$ (using $\sum_{n=0}^{-1} = 0$) and the optimal value function

$$V_N(x_0) = \inf_{u \in \mathcal{U}} J_N(x_0, u). \tag{2.7}$$

Note that this is the conceptually simplest receding horizon approach in which neither terminal costs nor terminal constraints are imposed.

Based on this finite horizon optimal value function, for $m \leq N$ we define an m-step feedback law $\mu_{N,m}$ by picking the first m elements of the optimal control sequence for this problem according to the following definition.

DEFINITION 2.3. *Let u^* be a minimizing control for* (2.6) *and initial value x_0. Then we define the m-step MPC feedback law by*

$$\mu_{N,m}(x_0, n) = u^*(n), \ \ n = 0, \dots, m-1.$$

Here the value N is called the optimization horizon *while we refer to m as the* control horizon.

Note that we do not need uniqueness of u^* for this definition; however, for $\mu_{N,m}(x_0, \cdot)$ being well defined we suppose that for each x_0 we select one specific u^* from the set of optimal controls.

The first goal of the present paper is to give estimates about the suboptimality of the feedback $\mu_{N,n}$ for the infinite horizon problem. More precisely, for an m-step feedback law μ with corresponding solution trajectory $x_\mu(n)$ from (2.4) we define

$$V_\infty^\mu(x_0) := \sum_{n=0}^{\infty} l(x_\mu(n), \mu(x_\mu([n]_m), n - [n]_m))$$

and are interested in upper bounds for the infinite horizon value $V_\infty^{\mu_{N,m}}$, i.e., in an estimate about the "degree of suboptimality" of the controller $\mu_{N,m}$. Based on this estimate, the second purpose of this paper is to derive results on the asymptotic stability of the resulting closed loop system using V_N as a Lyapunov function.

The approach we take in this paper relies on results on relaxed dynamic programming [16, 23] which were already used in an MPC context in [7, 8]. Next we state the basic relaxed dynamic programming inequality adapted to our setting.

PROPOSITION 2.4. *Consider an m-step feedback law $\tilde{\mu} : X \times \{0, \dots, m-1\} \to U$, the corresponding solution $x_{\tilde{\mu}}(k)$ with $x_{\tilde{\mu}}(0) = x_0$, and a function $\widetilde{V} : X \to \mathbb{R}_0^+$ satisfying the inequality*

$$\widetilde{V}(x_0) \geq \widetilde{V}(x_{\tilde{\mu}}(m)) + \alpha \sum_{k=0}^{m-1} l(x_{\tilde{\mu}}(k), \tilde{\mu}(x_0, k)) \tag{2.8}$$

for some $\alpha \in (0, 1]$ and all $x_0 \in X$. Then for all $x \in X$ the estimate

$$\alpha V_\infty(x) \leq \alpha V_\infty^{\tilde{\mu}}(x) \leq \widetilde{V}(x)$$

holds.

Proof. The proof is similar to that of [23, Proposition 3] and [8, Proposition 2.2]: Consider $x_0 \in X$ and the trajectory $x_{\tilde{\mu}}(n)$ generated by the closed loop system using $\tilde{\mu}$. Then from (2.8) for all $n \in \mathbb{N}_0$ we obtain

$$\alpha \sum_{k=0}^{m-1} l(x_{\tilde{\mu}}(nm + k), \tilde{\mu}(x_{\tilde{\mu}}(nm), k)) \leq \widetilde{V}(x_{\tilde{\mu}}(mn)) - \widetilde{V}(x_{\tilde{\mu}}(m(n+1))).$$

Summing over n yields

$$\alpha\sum_{n=0}^{Km} l(x_{\tilde\mu}(n),\tilde\mu(x_{\tilde\mu}(n),\tilde\mu(x_{\tilde\mu}([n]_m),n-[n]_m)) = \alpha\sum_{n=0}^{K}\sum_{k=0}^{m-1} l(x_{\tilde\mu}(nm+k),\tilde\mu(x_{\tilde\mu}(nm),k))$$
$$\le \widetilde V(x(0)) - \widetilde V(x(mK)) \le \widetilde V(x(0)).$$

For $K\to\infty$ this yields that $\widetilde V$ is an upper bound for $\alpha V_\infty^{\tilde\mu}$ and, hence,

$$\alpha V_\infty(x) \le \alpha V_\infty^{\tilde\mu}(x) \le \widetilde V(x). \qquad \square$$

Remark 2.5. The term "unconstrained" refers only to constraints which are introduced in order to ensure stability of the closed loop. Other constraints can be easily included in our setup; e.g., the set U of admissible control values could be subject to—possibly state dependent—constraints or X could be the feasible set of a state constrained problem on a larger state space.

3. Asymptotic controllability and optimal values. In this section we introduce an asymptotic controllability assumption and deduce several consequences for our optimal control problem. In order to facilitate this relation, we will formulate our basic controllability assumption, below, not in terms of the trajectory but in terms of the running cost l along a trajectory.

To this end we say that a continuous function $\rho:\mathbb{R}_{\ge0}\to\mathbb{R}_{\ge0}$ is of class $\mathcal{K}_\infty$ if it satisfies $\rho(0)=0$, and is strictly increasing and unbounded. We say that a continuous function $\beta:\mathbb{R}_{\ge0}\times\mathbb{R}_{\ge0}\to\mathbb{R}_{\ge0}$ is of class $\mathcal{KL}_0$ if for each $r>0$ we have $\lim_{t\to\infty}\beta(r,t)=0$ and for each $t\ge0$ we either have $\beta(\cdot,t)\in\mathcal{K}_\infty$ or $\beta(\cdot,t)\equiv 0$. Note that in order to allow for tighter bounds for the actual controllability behavior of the system we use a larger class than the usual class $\mathcal{KL}$. It is, however, easy to see that each $\beta\in\mathcal{KL}_0$ can be overbounded by a $\tilde\beta\in\mathcal{KL}$, e.g., by setting $\tilde\beta(r,t)=\max_{\tau\ge t}\beta(r,\tau)+e^{-t}r$. Furthermore, we define $l^*(x):=\min_{u\in U} l(x,u)$.

Assumption 3.1. Given a function $\beta\in\mathcal{KL}_0$, for each $x_0\in X$ there exists a control function $u_{x_0}\in\mathcal{U}$ satisfying

$$l(x(n,u_{x_0}),u_{x_0}(n)) \le \beta(l^*(x_0),n)$$

for all $n\in\mathbb{N}_0$. Special cases for $\beta\in\mathcal{KL}_0$ are

$$\beta(r,n)=C\sigma^n r \tag{3.1}$$

for real constants $C\ge1$ and $\sigma\in(0,1)$, i.e., *exponential controllability*, and

$$\beta(r,n)=c_n r \tag{3.2}$$

for some real sequence $(c_n)_{n\in\mathbb{N}_0}$ with $c_n\ge0$ and $c_n=0$ for all $n\ge n_0$, i.e., *finite time controllability* (with linear overshoot).

For certain results, it will be useful to have the property

$$\beta(r,n+m)\le\beta(\beta(r,n),m) \quad \text{for all } r\ge0, n,m\in\mathbb{N}_0. \tag{3.3}$$

Property (3.3) ensures that any sequence of the form $\lambda_n=\beta(r,n)$, $r>0$, also fulfills $\lambda_{n+m}\le\beta(\lambda_n,m)$. It is, for instance, always satisfied in case (3.1) and satisfied in case (3.2) if $c_{n+m}\le c_nc_m$. If needed, this property can be assumed without loss of generality, because by Sontag's $\mathcal{KL}$-Lemma [25] β in Assumption 3.1 can be replaced by a β of the form $\beta(r,t) = \alpha_1(\alpha_2(r)e^{-t})$ for $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$. Then, (3.3) is easily verified if $\alpha_2 \circ \alpha_1(r) \geq r$ which is equivalent to $\alpha_1 \circ \alpha_2(r) \geq r$, which in turn is a necessary condition for Assumption 3.1 to hold for $n = 0$ and $\beta(r,t) = \alpha_1(\alpha_2(r)e^{-t})$.

Remark 3.2. Computing β satisfying Assumption 3.1 is in general a hard task for nonlinear systems. One way to obtain such a β is via a suitable control Lyapunov function, similar to the procedure described in [15, section 4.4] or used in [20, Proof of Proposition 1]. However, as we will see later, the precise knowledge of β is not necessarily needed in order to apply our results, because we will be able to identify structural properties of β which guarantee good performence of the MPC closed loop; cf. the design guidelines and the PDE example (7.2) in section 7.

Under Assumption 3.1, for any $r \geq 0$ and any $N \geq 1$ we define the value

$$B_N(r) := \sum_{n=0}^{N-1} \beta(r,n). \tag{3.4}$$

An immediate consequence of Assumption 3.1 is the following lemma.

LEMMA 3.3. *For each $N \geq 1$ the inequality*

$$V_N(x_0) \leq B_N(l^*(x_0)) \tag{3.5}$$

holds.

Proof. Using u_{x_0} from Assumption 3.1, the inequality follows immediately from

$$\begin{aligned} V_N(x_0) \leq J_N(x_0, u_{x_0}) &= \sum_{n=0}^{N-1} l(x(n, u_{x_0}), u_{x_0}(n)) \\ \leq \sum_{n=0}^{N-1} \beta(l^*(x_0), n) &= B_N(l^*(x_0)). \qquad \square \end{aligned}$$

In the special case (3.1) B_N, $N \geq 1$, evaluates to

$$B_N(r) = C\frac{1-\lambda^N}{1-\lambda}r,$$

while for (3.2) we obtain

$$B_N(r) = C_N r, \text{ where } C_N = \sum_{j=0}^{\min\{n_0, N-1\}} c_n.$$

The following lemma gives bounds on the finite horizon functional along optimal trajectories. It uses the fact that final pieces of obtimal trajectories are again optimal trajectories to which we can apply Lemma 3.3.

LEMMA 3.4. *Assume Assumption* 3.1 *and consider $x_0 \in X$ and an optimal control u^* for the finite horizon optimal control problem* (2.7) *with optimization horizon $N \geq 1$. Then for each $k = 0, \ldots, N-1$ the inequality*

$$J_{N-k}(x_{u^*}(k), u^*(k+\cdot)) \leq B_{N-k}(l^*(x_{u^*}(k))$$

holds for B_N from (3.4).

Proof. Pick any $k \in \{0, \ldots, N-1\}$. Using u_{x_0} from Assumption 3.1 with $x_0 = x_{u^*}(k)$, from (3.5) we obtain

$$J_{N-k}(x_{u^*}(k), u_{x_0}(\cdot)) \leq B_{N-k}(l^*(x_{u^*}(k))). \tag{3.6}$$

Hence, for the control function defined by

$$\tilde{u}(n) = \begin{cases} u^*(n), & n \le k-1 \\ u_{x_0}(n), & n \ge k \end{cases}$$

we obtain

$$V_N(x_0) \le J_N(x_0, \tilde{u}) = J_k(x_0, u^*) + J_{N-k}(x_{u^*}(k), u_{x_0}(\cdot)).$$

On the other hand, we have

$$V_N(x_0) = J_N(x_0, u^*) = J_k(x_0, u^*) + J_{N-k}(x_{u^*}(k), u^*(k+\cdot)).$$

Subtracting the latter from the former yields

$$0 \le J_{N-k}(x_{u^*}(k), u_{x_0}(\cdot)) - J_{N-k}(x_{u^*}(k), u^*(k+\cdot)),$$

which using (3.6) implies

$$J_{N-k}(x_{u^*}(k), u^*(k+\cdot)) \le J_{N-k}(x_{u^*}(k), u_{x_0}(\cdot)) \le B_{N-k}(l^*(x_{u^*}(k)),$$

i.e., the assertion. □

A similar inequality can be obtained for V_N. Here we split up a trajectory into two pieces and apply Lemma 3.3 to the second piece.

LEMMA 3.5. *Assume Assumption* 3.1 *and consider* $x_0 \in X$ *and an optimal control* u^* *for the finite horizon optimal control problem* (2.7) *with optimization horizon* N. *Then for each* $m = 1, \dots, N-1$ *and each* $j = 0, \dots, N-m-1$ *the inequality*

$$V_N(x_{u^*}(m)) \le J_j(x_{u^*}(m), u^*(m+\cdot)) + B_{N-j}(l^*(x_{u^*}(m+j))$$

holds for B_N *from* (3.4).

Proof. We define the control function

$$\tilde{u}(n) = \begin{cases} u^*(m+n), & n \le j-1 \\ u_{x_0}(n), & n \ge j \end{cases}$$

for u_{x_0} from Assumption 3.1 with $x_0 = x_{u^*}(m+j)$. Then we obtain

$$\begin{aligned} V_N(x_{u^*}(m)) &\le J(x_{u^*}(m), \tilde{u}) \\ &= J_j(x_{u^*}(m), u^*(m+\cdot)) + J_{N-j}(x_{u^*}(m+j), u_{x_0}) \\ &\le J_j(x_{u^*}(m), u^*(m+\cdot)) + B_{N-j}(l^*(x_{u^*}(m+j))) \end{aligned}$$

where we used (3.5) in the last step. This is the desired inequality. □

4. Computation of performance bounds. In this section we provide a constructive approach in order to compute α in (2.8) for systems satisfying Assumption 3.1. For this purpose we consider arbitrary values $\lambda_0, \dots, \lambda_{N-1} > 0$ and $\nu > 0$ and start by deriving necessary conditions under which these values coincide with an optimal sequence $l(x_{u^*}(n), u^*(n))$ and an optimal value $V_N(x_{u^*}(m))$, respectively.

PROPOSITION 4.1. *Assume Assumption* 3.1 *and consider* $N \ge 1$, $m \in \{1, \dots, N-1\}$, *a sequence* $\lambda_n > 0$, $n = 0, \dots, N-1$ *a value* $\nu > 0$. *Consider* $x_0 \in X$ *and assume that there exists an optimal control function* $u^* \in \mathcal{U}$ *for the finite horizon problem* (2.7) *with horizon length* N, *such that*

$$\lambda_n = l(x_{u^*}(n), u^*(n)), \quad n = 0, \dots, N-1$$

holds. Then

$$\sum_{n=k}^{N-1} \lambda_n \le B_{N-k}(\lambda_k), \quad k = 0, \dots, N-2 \tag{4.1}$$

holds. If, furthermore,

$$\nu = V_N(x_{u^*}(m)),$$

holds, then

$$\nu \le \sum_{n=0}^{j-1} \lambda_{n+m} + B_{N-j}(\lambda_{j+m}), \quad j = 0, \dots, N-m-1 \tag{4.2}$$

holds.

Proof. If the stated conditions hold, then λ_n and ν must meet the inequalities given in Lemmas 3.4 and 3.5, which is exactly (4.1) and (4.2). □

Using this proposition we can give a sufficient condition for suboptimality of the MPC feedback law $\mu_{N,m}$. The idea behind the following theorem is to express the terms in inequality (2.8) using the values $\lambda_0, \dots, \lambda_{N-1}$ and ν introduced above.

THEOREM 4.2. *Consider $\beta \in \mathcal{KL}_0$, $N \ge 1$, $m \in \{1, \dots, N-1\}$, and assume that all sequences $\lambda_n > 0$, $n = 0, \dots, N-1$ and values $\nu > 0$ fulfilling* (4.1), (4.2) *satisfy the inequality*

$$\sum_{n=0}^{N-1} \lambda_n - \nu \ge \alpha \sum_{n=0}^{m-1} \lambda_n \tag{4.3}$$

for some $\alpha \in (0, 1]$.

Then for each optimal control problem (2.1), (2.7) *satisfying Assumption* 3.1, *the assumptions of Proposition* 2.4 *are satisfied for the m-step MPC feedback law $\mu_{N,m}$ and, in particular, the inequality*

$$\alpha V_\infty(x) \le \alpha V_\infty^{\mu_{N,m}}(x) \le V_N(x)$$

holds for all $x \in X$.

Proof. Consider an initial value $x_0 \in X$ and the m-step MPC-feedback law $\mu_{N,m}$. Then there exists an optimal control u^* for x_0 such that

$$u^*(k) = \mu_{N,m}(x_0, k), \;\; k = 0, \dots, m-1 \quad \text{and} \quad x_{\mu_{N,m}}(k) = x_{u^*}(k), \;\; k = 0, \dots, m$$

and, consequently, also

$$l(x_{\mu_{N,m}}(k), \mu_{N,m}(x_0, k)) = l(x_{u^*}(k), u^*(k)), \quad k = 0, \dots, m-1$$

holds. These equalities imply

$$V_N(x_{\mu_{N,m}}(m)) + \alpha \sum_{n=0}^{m-1} l(x_{\mu_{N,m}}(n), \mu_{N,m}(x_0, n)) = V_N(x_{u^*}(m)) + \alpha \sum_{n=0}^{m-1} l(x_{u^*}(n), u^*(n)) \tag{4.4}$$

for any $\alpha \in \mathbb{R}$.

Now by Proposition 4.1 the values $\lambda_n = l(x_{u^*}(k), u^*(k))$ and $\nu = V_N(x_{u^*}(m))$ satisfy (4.1) and (4.2), and hence, by assumption also (4.3). Thus, we obtain

$$\begin{aligned} V_N(x_{u^*}(m)) + \alpha \sum_{n=0}^{m-1} l(x_{u^*}(n), u^*(n)) &= \nu + \alpha \sum_{n=0}^{m-1} \lambda_n \;\leq\; \sum_{n=0}^{N-1} \lambda_n \\ &= \sum_{n=0}^{N-1} l(x_{u^*}(n), u^*(n)) \;=\; V_N(x_0). \end{aligned}$$

Together with (4.4) this yields (2.8) and, thus, the assertion. □

Remark 4.3. Our analysis is easily extended to more general settings. As an example, we show how an additional weight on the final term in the finite horizon optimal control problem can be included. In this case, the functional J_N is generalized to

$$J_N^\omega(x_0, u) = \sum_{n=0}^{N-2} l(x_u(n), u(n)) + \omega l(x_u(N-1), u(N-1)) \tag{4.5}$$

for some $\omega \geq 1$. Note that the original form of the functional J_N from (2.6) is obtained by setting $\omega = 1$, i.e., $J_N = J_N^1$. A straightforward extension of the proofs in the previous section reveals that the inequalities in Lemma 3.4 and Lemma 3.5 become

$$J_{N-k}^\omega(x_{u^*}(k), u^*(k+\cdot)) \leq B_{N-k}^\omega(l^*(x_{u^*}(k))$$

and

$$V_N(x_{u^*}(m)) \leq J_j^1(x_{u^*}, u^*(m+\cdot)) + B_{N-j}^\omega(l^*(x_{u^*}(m+j))),$$

respectively, with

$$B_N^\omega(r) := \sum_{n=0}^{N-2} \beta(r,n) + \omega\beta(r, N-1).$$

Consequently, the inequalities (4.1), (4.2), and (4.3) change to

$$\sum_{n=k}^{N-2} \lambda_n + \omega\lambda_{N-1} \leq B_{N-k}^\omega(\lambda_k), \qquad \nu \leq \sum_{n=0}^{j-1} \lambda_{n+m} + B_{N-j}^\omega(\lambda_{j+m})$$

and

$$\sum_{n=0}^{N-2} \lambda_n + \omega\lambda_{N-1} - \nu \geq \alpha \sum_{n=0}^{m-1} \lambda_n,$$

respectively.

In view of Theorem 4.2, the value α can be interpreted as a performance bound which indicates how good the receding horizon MPC strategy approximates the infinite horizon problem. In the remainder of this section we present an optimization approach for computing α. To this end consider the following optimization problem.

PROBLEM 4.4. Given $\beta \in \mathcal{KL}_0$, $N \geq 1$ and $m \in \{1, \ldots, N-1\}$, compute

$$\alpha := \inf_{\lambda_0, \ldots, \lambda_{N-1}, \nu} \frac{\sum_{n=0}^{N-1} \lambda_n - \nu}{\sum_{n=0}^{m-1} \lambda_n}$$

subject to the constraints (4.1) and (4.2) and

$$\lambda_0, \ldots, \lambda_{N-1}, \nu > 0. \tag{4.6}$$

The following is a straightforward corollary from Theorem 4.2.

COROLLARY 4.5. *Consider $\beta \in \mathcal{KL}_0$, $N \geq 1$, $m \in \{1, \ldots, N-1\}$, and assume that the optimization Problem* 4.4 *has an optimal value $\alpha \in (0,1]$.*

Then for each optimal control problem (2.1), (2.7) *satisfying Assumption* 3.1, *the assumptions of Proposition* 2.4 *are satisfied for the m-step MPC feedback law $\mu_{N,m}$ and, in particular, the inequality*

$$\alpha V_\infty(x) \leq \alpha V_\infty^{\mu_{N,m}}(x) \leq V_N(x)$$

holds for all $x \in X$.

Proof. The proof follows immediately from Theorem 4.2 and the definition of Problem 4.4. □

Problem 4.4 is an optimization problem of a much lower complexity than the original MPC optimization problem. Still, it is in general nonlinear. However, it becomes a linear program if we assume that $\beta(r,n)$ and, thus, $B_k(r)$ are linear in r.

LEMMA 4.6. *If $\beta(r,t)$ is linear in r, then Problem* 4.4 *yields the same optimal value α as*

$$\alpha := \min_{\lambda_0, \lambda_1, \ldots, \lambda_{N-1}, \nu} \sum_{n=0}^{N-1} \lambda_n - \nu \tag{4.7}$$

subject to the (now linear) constraints (4.1) *and* (4.2) *and*

$$\lambda_0, \ldots, \lambda_{N-1}, \nu \geq 0, \quad \sum_{n=0}^{m-1} \lambda_n = 1. \tag{4.8}$$

Proof. Due to the linearity, all sequences $\bar\lambda_0, \ldots, \bar\lambda_{N-1}, \bar\nu$ satisfying (4.1), (4.2), and (4.6) can be written as $\gamma\lambda_0, \ldots, \gamma\lambda_{N-1}, \gamma\nu$ for some $\lambda_0, \ldots, \lambda_{N-1}, \nu$ satisfying (4.1), (4.2), (4.6), and (4.8), where $\gamma = (\sum_{n=0}^{m-1} \bar\lambda_n)^{-1}$. Since

$$\frac{\sum_{n=0}^{N-1} \bar\lambda_n - \bar\nu}{\sum_{n=0}^{m-1} \bar\lambda_n} = \frac{\sum_{n=0}^{N-1} \gamma\lambda_n - \gamma\nu}{\sum_{n=0}^{m-1} \gamma\lambda_n} = \frac{\sum_{n=0}^{N-1} \lambda_n - \nu}{\sum_{n=0}^{m-1} \lambda_n} = \sum_{n=0}^{N-1} \lambda_n - \nu,$$

under the constraints (4.6) and (4.8) the values α in Problem 4.4 and (4.7) coincide. Now by continuity we can weaken (4.6) to $\lambda_1, \ldots, \lambda_{N-1}, \nu \geq 0$ without changing α in (4.7); i.e., we can omit the constraints (4.6) in the linear problem. This shows the claim. □

MATLAB implementations for the linear program described in Lemma 4.6 for (3.1) and (3.2), including also the weights ω from Remark 4.3, are available from the web site: www.math.uni-bayreuth.de/∼lgruene/publ/mpcbound.html.

Remark 4.7. Although restrictive, the linearity condition in Lemma 4.6 appears to be a natural condition for proving asymptotic stability of the MPC closed loop. In fact, in many papers in the MPC literature for nonlinear finite dimensional systems (e.g., in [3], [17], [13]) stabilizability or even controllability of the linearization at the equilibrium is assumed. In our framework this implies exponential controllability on any compact subset of the state space for any quadratic cost function and, thus, linearity of $\beta(r,t)$ in r. In the more general approach in [4], asymptotic stability (in contrast to mere *practical* asymptotic stability) was proved under a condition which implies

$$l(x(n,u_{x_0})) \leq Ce^{-\sigma n} l(x_0)$$

for some suitable function l, cf. [4, Corollary 2], which yields our exponential controllability assumption for this l and, thus, again linearity of $\beta(r,t)$ in r.

5. Asymptotic stability. In this section we show how the performance bound α can be used in order to conclude asymptotic stability of the MPC closed loop. More precisely, we investigate the asymptotic stability of the zero set of l^*. To this end we make the following assumption.

Assumption 5.1. There exists a closed set $A \subset X$ satisfying:

(i) For each $x \in A$ there exists $u \in U$ with $f(x,u) \in A$ and $l(x,u) = 0$; i.e., we can stay inside A forever at zero cost.

(ii) There exist $\mathcal{K}_\infty$–functions α_1, α_2 such that the inequality

$$\alpha_1(\|x\|_A) \leq l^*(x) \leq \alpha_2(\|x\|_A) \tag{5.1}$$

holds for each $x \in X$ where $\|x\|_A := \min_{y\in A} \|x-y\|$.

This assumption assures global asymptotic stability of A under the optimal feedback (2.5) for the infinite horizon problem, provided $\beta(r,n)$ is summable. We remark that condition (ii) can be relaxed in various ways, e.g., it could be replaced by a detectability condition similar to the one used in [4]. However, in order to keep the presentation in this paper technically simple we will work with Assumption 5.1(ii) here. Our main stability result is formulated in the following theorem. As usual, we say that a feedback law μ asymptotically stabilizes a set A if there exists $\tilde{\beta} \in \mathcal{KL}_0$ such that the closed loop system satisfies $\|x_\mu(n)\|_A \leq \tilde{\beta}(\|x_0\|_A, n)$.

THEOREM 5.2. *Consider $\beta \in \mathcal{KL}_0$, $N \geq 1$, $m \in \{1,\dots,N-1\}$, and assume that the optimization Problem* 4.4 *has an optimal value $\alpha \in (0,1]$.*

Then for each optimal control problem (2.1), (2.7) *satisfying the Assumptions* 3.1 *and* 5.1, *the m-step MPC feedback law $\mu_{N,m}$ asymptotically stabilizes the set A. Furthermore, V_N is a corresponding m-step Lyapunov function in the sense that*

$$V_N(x_{\mu_{N,m}}(m)) \leq V_N(x) - \alpha V_m(x). \tag{5.2}$$

Proof. In order to prove the theorem, we first show that $V_N(x_{\mu_{N,m}}(km))$ is decreasing to 0 as $k \to \infty$. In the second step we show that $V_N(x_{\mu_{N,m}}(n))$ is suitably bounded also for those times n which are not integer multiples of m.

From (5.1) and Lemma 3.3 we immediately obtain the inequality

$$\alpha_1(\|x\|_A) \leq V_N(x) \leq B_N(\alpha_2(\|x\|_A)). \tag{5.3}$$

Note that $B_N \circ \alpha_2$ is again a $\mathcal{K}_\infty$-function. The stated Lyapunov inequality (5.2) follows immediately from (2.8) which holds according to Corollary 4.5. Again using (5.1) we obtain $V_m(x) \geq \alpha_1(\|x\|_A)$ and, thus, a standard construction (see, e.g., [20, Proof of Proposition 1] or [15, section 4.4]) yields a $\mathcal{KL}$–function ρ for which the inequality

$$V_N(x_{\mu_{N,m}}(km)) \leq \rho(V_N(x), k)$$

holds. In addition, using the definition of $\mu_{N,m}$, for $n = 1, \dots, m-1$ we obtain

$$\begin{aligned}V_N(x_{\mu_{N,m}}(n)) &= \sum_{k=n}^{m-1} l(x_{\mu_{N,m}}(k), \mu_{N,m}(x_{\mu_{N,m}}(0), k)) + V_{N-m+n}(x_{\mu_{N,m}}(m)) \\ &\leq \sum_{k=0}^{m-1} l(x_{\mu_{N,m}}(k), \mu_{N,m}(x_{\mu_{N,m}}(0), k)) + V_{N-m+n}(x_{\mu_{N,m}}(m)) \\ &\leq V_N(x) + V_N(x_{\mu_{N,m}}(m)) \ \leq \ 2V_N(x),\end{aligned}$$

where we have used (5.2) in the last inequality. Thus, for all $n \in \mathbb{N}_0$ we obtain the estimate

$$V_N(x_{\mu_{N,m}}(n)) \leq 2\rho(V_N(x), [n]_m/m),$$

which eventually implies

$$\begin{aligned}\|x_{\mu_{N,m}}(n)\|_A &\leq \alpha_1^{-1}(V_N(x_{\mu_{N,m}}(n))) \ \leq \ \alpha_1^{-1}(2\rho(V_N(x), [n]_m/m)) \\ &\leq \alpha_1^{-1}(2\rho(B_N(\alpha_2(\|x\|_A)), [n]_m/m))\end{aligned}$$

and, thus, the desired asymptotic stability with $\mathcal{KL}$-function given by, e.g.,

$$\tilde{\beta}(r, n) = \alpha_1^{-1}(2\rho(B_N(\alpha_2(r)), [n]_m/m)) + re^{-n}. \qquad \square$$

Of course, Theorem 5.2 gives a conservative criterion in the sense that for a given system satisfying the Assumptions 3.1 and 5.1 asymptotic stability of the closed loop may well hold for smaller optimization horizons N. A trivial example for this is an asymptotically stable system (2.1) which does not depend on u at all, which will of course be "stabilized" regardless of N.

Hence, the best we can expect is that our condition is tight under the information we use; i.e., that given β, N, m such that the assumption of Theorem 5.2 is *violated* we can always find a system satisfying Assumptions 3.1 and 5.1 which is *not* stabilized by the MPC feedback law. The following Theorem 5.3 shows that this is, indeed, the case if β satisfies (3.3). Its proof relies on the explicit construction of a control system and a running cost for which the MPC closed loop is not asymptotically stable. Although this is in principle possible for all $m \in \{1, \dots, N-1\}$, we restrict ourselves to the classical feedback case, i.e., $m = 1$, in order to keep the construction technically simple.

THEOREM 5.3. *Consider $\beta \in \mathcal{KL}_0$ satisfying* (3.3)*, $N \geq 1$, $m = 1$ and assume that the optimization Problem* 4.4 *has an optimal value $\alpha < 0$.*

Then there exists an optimal control problem (2.1)*,* (2.7) *satisfying the Assumptions* 3.1 *and* 5.1 *which is not asymptotically stabilized by the MPC feedback law $\mu_{N,1}$.*

Proof. If $\alpha < 0$, then there exists $\lambda_n, \nu > 0$ meeting the constraints of Problem 4.4 satisfying $\sum_{n=0}^{N-1} \lambda_n - \nu/(\sum_{n=0}^{m-1} \lambda_n) =: \tilde{\alpha} < 0$. By Lemma 9.1 we can without loss of generality assume that the inequalities (4.1) are strict for λ_n.

Now we construct an optimal control problem on the set $X = \{0\} \cup \{2^{-k} | k \in \mathbb{N}_0\} \times \{-N+1, \dots, N\}$ with control values $U = \{-1, 0, 1\}$ and dynamics given by

$$\begin{aligned}f((1,p),-1) &= (1, \max\{-N+1, p-1\}) \\ f((1,p),0) &= (1/2, p) \\ f((1,p),1) &= (1, \min\{N, p+1\}) \\ f((q,p),u) &= (q/2, p), \qquad\qquad q \leq 1/2, u \in U.\end{aligned}$$

The running cost is given by

$$\begin{aligned} l((1,p),1) &= \lambda_p, && p \in \{0, N-1\}\\ l((1,p),1) &= \nu, && p \notin \{0, N-1\}\\ l((1,p),-1) &= l((1,-p+1),1) &&\\ l((1,p),0) &= \beta(\min\{l((1,n),1), l((1,n),-1)\},0) &&\\ l((2^{-k},p),u) &= \beta(\min\{l((1,p),1), l((1,p),-1)\},k), && k \geq 1,\, u \in U \end{aligned}$$

We intend to show that the set $A = \{x \in X \,|\, l^*(x) = 0\}$ is not asymptotically stabilized. This set A satisfies Assumption 5.1(i) for $u = 0$ and (ii) for $\tilde{\alpha}_1(r) = \inf_{x\in X, \|x\|_A \geq r} l^*(x)$ and $\tilde{\alpha}_2(r) = \sup_{x\in X, \|x\|_A \leq r} l^*(x)$. Due to the discrete nature of the state space $\tilde{\alpha}_1$ and $\tilde{\alpha}_2$ are discontinuous but they are easily under- and overbounded by continuous $\mathcal{K}_\infty$ functions α_1 and α_2, respectively. Furthermore, by virtue of (3.3) the optimal control problem satisfies Assumption 3.1 for $u_x \equiv 0$.

Now we prove the existence of a trajectory which does not converge to A, which shows that asymptotic stability does not hold. To this end we abbreviate $\Lambda = \sum_{n=0}^{N-1} \lambda_n$ (note that (9.1) implies $\nu > \lambda$) and investigate the values $J_N((1,0),u)$ for different choices of u:

Case 1: $u(0) = 0$. In this case, regardless of the values $u(n)$, $n \geq 1$, we obtain $x(n,u) = (2^{-n}, 0)$ and, thus,

$$\begin{aligned} J_N((1,0),u) &= \sum_{n=0}^{N-1} \beta(\min\{l((1,0),1), l((1,0),-1)\}, n)\\ &= B_N(\min\{l((1,0),1), l((1,0),-1)\}) \;=\; B_N(\min\{\lambda_0, \lambda_1\}). \end{aligned}$$

In case that the minimum is attained in λ_0 by the (strict) inequality (4.1) for $k = 0$ we obtain $J_N((1,0),u) > \Lambda$. If the minimum is attained in λ_1, then by (4.2) for $j = 0$ and (9.1) we obtain $J_N((1,0),u) \geq \nu > \Lambda$. Thus, in both cases the inequality $J_N((1,0),u) > \Lambda$ holds.

Case 2: $u(n) = -1$, $n = 0, \dots, N-2$. This choice yields $x(n,u) = (1,-n)$ for $n = 0, \dots, N-2$ and, thus,

$$J_N((1,0),u) = \sum_{n=0}^{N-2} \lambda_{n+1} + l((1,-N+1), u(N-1)) \geq l((1,-N+1), u(N-1)) \geq \nu > \Lambda.$$

Case 3: $u(n) = -1$, $n = 0, \dots, k-1$, and $u(k) = 1$ for a $k \in \{1, \dots, N-2\}$. In this case we obtain $x(n,u) = (1,-n)$ for $n = 0, \dots, k$ implying

$$J_N((1,0),u) = \sum_{n=0}^{k-1} \lambda_{n+1} + l((1,-k),1) \geq l((1,-k),1) = \nu > \Lambda.$$

Case 4: $u(n) = -1$, $n = 0, \dots, k-1$, and $u(k) = 0$ for a $k \in \{1, \dots, N-2\}$. This control sequence yields $x(n,u) = (1,-n)$ for $n = 0, \dots, k$ and $x(n,u) = (2^{-(n-k)}, -k)$ for $n = k+1, \dots, N-1$ and, thus,

$$\begin{aligned} J_N((1,0),u) &= \sum_{n=0}^{k-1} \lambda_{n+1} + \sum_{n=k}^{N-1} \beta(\min\{l((1,-k),1), l((1,-k),-1)\}, n-k)\\ &= \sum_{n=0}^{k-1} \lambda_{n+1} + B_{N-k}(\lambda_{k+1}) \geq \nu > \Lambda, \end{aligned}$$

where we have used (4.2) for $j = k$ in the second last inequality.

Case 5: $u(n) = 1$, $n = 0, \dots, N-1$. This yields $x(n,u) = (1,n)$ and, thus,

$$J_N((1,0),u) = \sum_{n=0}^{N-1} \lambda_n = \Lambda.$$

Summarizing, we obtain that any optimal control u_x^* for $x = (1,0)$ must satisfy $u_x^*(0) = 1$ because for $u(0) = 1$ we can realize a value $\leq \Lambda$, while for $u(0) \neq 1$ we inevitably obtain a value $> \Lambda$. Consequently, the MPC feedback law will steer the system from $x = (1,0)$ to $x^+ := (1,1)$.

Now we use that by construction f and l have the symmetry properties

$$f((q,p),u) - (0,p) = -f((q,-p+1),-u) + (0,-p+1), \quad l((q,p),u) = l((q,-p+1),-u)$$

for all $(q,p) \in X$ which implies $J((q,p),u) = J(q,-p+1),-u)$. Observe that $x^+ = (1,1)$ is exactly the symmetric counterpart of $x = (1,0)$. Thus, any optimal control $u_{x^+}^*$ from x^+ must satisfy $u_{x^+}^*(n) = -u_x^*(n)$ for some optimal control u_x^* for initial value x. Hence, we obtain $u_{x^+}^*(0) = -1$ which means that the MPC feedback steers x^+ back to x. Thus, under the MPC-feedback law we obtain the closed loop trajectory $(x, x^+, x, x^+, \dots)$ which clearly does not converge to A. This shows that the closed loop system is not asymptotically stable. □

6. Analysis of MPC schemes.

Using the optimization Problem 4.4 we are now able to analyze the optimization horizon N needed in order to ensure stability and desired performance of the MPC closed loop. More precisely, given β from Assumption 3.1 and a desired $\alpha_0 \geq 0$, by solving Problem 4.4 we can compute the minimal horizon

$$\widehat{N} := \min\{N \in \mathbb{N} \mid \alpha > \alpha_0\} \tag{6.1}$$

which yields asymptotic stability and—in case $\alpha_0 > 0$—ensures the performance

$$V_\infty^{\mu_{\widehat{N},m}}(x) \leq V_{\widehat{N}}(x)/\alpha_0.$$

Note that even without sophisticated algorithms for finding the minimum in (6.1) the determination of $\widehat{N}$ needs at most a couple of seconds using our MATLAB code.

We first observe that α from Problem 4.4 is monotone decreasing in β, i.e., for β_1 and $\beta_2 \in \mathcal{KL}_0$ satisfying $\beta_1(r,n) \geq \beta_2(r,n)$ for all $r \in \mathbb{R}_{\geq 0}$, $n \in \mathbb{N}_0$, we obtain $\alpha_1 \leq \alpha_2$ for the corresponding solutions of Problem 4.4. This property immediately follows from the fact that a smaller β induces stronger constraints in the optimization problem. Consequently, the horizon $\widehat{N}$ in (6.1) is monotone increasing in β. We emphasize that this is an important feature because in practice it will rarely be possible to compute a tight bound β in Assumption 3.1 and typically only a—more or less—conservative upper bound will be available. Then the monotonocity property ensures that any $\widehat{N}$ computed using such an upper bound β will also give an upper bound on the real minimal horizon $\widehat{N}$ for the system.

In the sequel, we will on the one hand investigate how different choices of the control horizon m and the terminal weight ω (cf. Remark 4.3) affect the horizon N. On the other hand, we will highlight how different characteristic features of β in Assumption 3.1, e.g., overshoot and decay rate, influence the horizon $\widehat{N}$. Since the controllability Assumption 3.1 involves the running cost l, the results of this latter analysis will, in particular, yield guidelines for the choice of l allowing to design stable

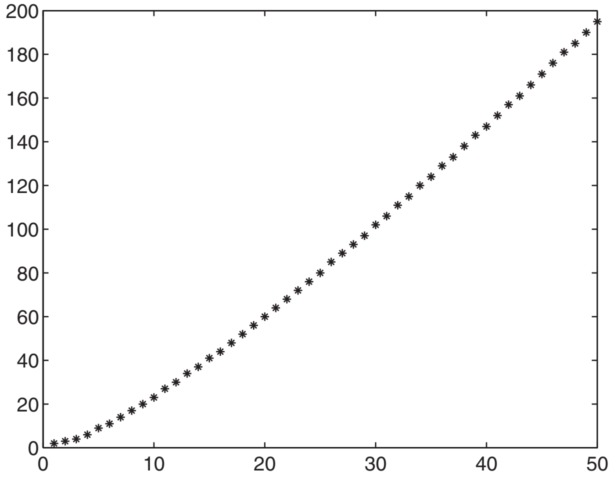


FIG. 6.1. *Minimal stabilizing horizon $\widehat{N}$ for $m = 1$.*

MPC schemes with small optimization horizons, which we formulate and illustrate in the ensuing section 7 for finite and infinite dimensional examples. In our analysis we will concentrate on mere asymptotic stability; i.e., we will consider $\alpha_0 = 0$, however, all computations yield qualitatively similar results for $\alpha_0 > 0$. In what follows, for the sake of brevity we concentrate on a couple of particularly illuminating controllability functions β, noting that much more details could be investigated, if desired.

We start by investigating how our estimated minimal stabilizing horizon N depends on the accumulated overshoot represented by β, i.e., on the value $\gamma > 0$ satisfying

$$\sum_{n=0}^{\infty} \beta(r,n) \leq \gamma r. \tag{6.2}$$

To this end, we use the observation that if N is large enough in order to stabilize each system satisfying Assumption 3.1 with

$$\beta(r,0) = \gamma r, \;\; \beta(r,n) = 0, \, n \geq 1, \tag{6.3}$$

then N is also large enough to stabilize each system satisfying Assumption 3.1 with β from (6.2). In particular, this applies to $\beta(r,n) = C\sigma^n r$ with $C/(1-\sigma) \leq \gamma$. The reason for this is that the inequalities (4.1), (4.2) for (6.3) form weaker constraints than the respective inequalities for (6.2); hence, the minimal value α for (6.3) must be less or equal than α for (6.2).

Thus, we investigate the "worst case" (6.3) numerically and compute how the minimal stabilizing N depends on γ. To this end we computed $\widehat{N}$ from (6.1) for β from (6.3) with $\gamma = 1, 2, \ldots, 50$ and $m = 1$. The resulting values $\widehat{N}$ are shown in Figure 6.1.

It is interesting to observe that the resulting values almost exactly satisfy $\widehat{N} \approx \gamma \log \gamma$, which leads to the conjecture that this expression describes the analytical "stability margin".

In order to see the influence of the control horizon m, we have repeated this computation for $m = [N/2] + 1$, which numerically appears to be the optimal choice of m. The results are shown in Figure 6.2.

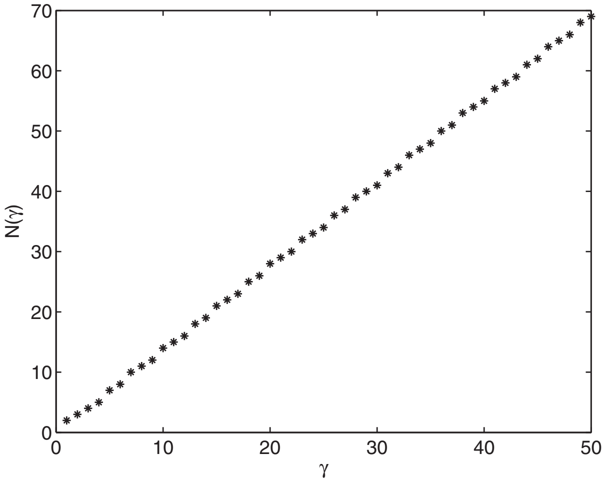


FIG. 6.2. *Minimal stabilizing horizon $\widehat{N}$ for $m = [N/2] + 1$.*

Here, one numerically observes $\widehat{N} \approx 1.4\gamma$; i.e., we obtain a linear dependence between γ and $\widehat{N}$ and, in particular, we obtain stability for much smaller N than in the case $m = 1$. However, when using such control horizons $m > 1$, one should keep in mind that the control loop is closed only every m steps, i.e., the re-computation of the control value based on the current measurement is performed at the times $0, m, 2m, \ldots$. This implies that the larger m is chosen, the more limited the ability of the feedback controller to react to perturbations (caused, e.g., by external disturbances or modelling errors) becomes. On the other hand, if a large overshoot γ cannot be avoided and hardware constraints restrict the computational resources, then moderately increasing m may provide a good compromise in order to reduce N and, thus, the complexity of the optimization problem to be solved online.

Figures 6.1 and 6.2 show how fast the necessary control horizon grows depending on γ, and obviously the smaller γ is, the smaller $\widehat{N}$ becomes. Note that $\gamma+1$ coincides with the parameter γ used in [8]. However, while the analysis in [8] is restricted to investigating the effect of this parameter γ, our optimization-based approach now allows for a detailed analysis on how $\widehat{N}$ varies for different functions β leading to the same value γ. For instance, in an exponentially decaying running cost with $\beta(r,n) = C\sigma^n r$, it will be interesting to know whether small overshoot (i.e., small C) or fast decay (i.e., small σ) are more important in order to ensure stability for small $\widehat{N}$. In order to analyze this dependence, we consider the classical feedback case $m = 1$ and compare the four different functions of the form $\beta(r,n) = C\sigma^n r$ with

$$\begin{array}{lllllll} & \text{(a)} & C = 3, & \sigma = 1/2 & \text{(b)} & C = 12/5, & \sigma = 3/5 \\ (6.4) & & & & & & \\ & \text{(c)} & C = 3/2, & \sigma = 3/4 & \text{(d)} & C = 6/5, & \sigma = 4/5. \end{array}$$

These four functions have in common that $\gamma = C/(1-\sigma) = 6$, but—as illustrated in Figure 6.3 for $r = 1$—they differ in both the size of the overshoot C, which is decreasing from (a) to (d) and the speed of decay σ which becomes slower from (a) to (d).

It is surprising to see how much the minimal stabilizing horizons $\widehat{N}$ differ from (a) to (d): solving (6.1) using Problem 4.4 we obtain (a) $\widehat{N} = 11$, (b) $\widehat{N} = 10$, (c) $\widehat{N} = 7$, and (d) $\widehat{N} = 4$. Thus, in order to ensure stability with small optimization horizon N

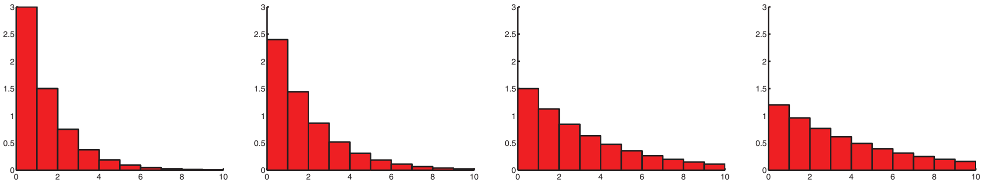


FIG. 6.3. *Exponentially decaying functions β with C, σ from* (6.4) (a)–(d) *(left to right).*

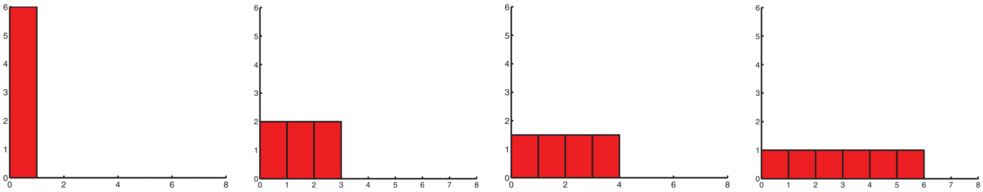


FIG. 6.4. *Finite time decaying functions β from* (6.5) (a)–(d) *(left to right).*

for exponentially decaying β in Assumption 3.1, small overshoot is considerably more important than fast decay.

A similar analysis can be carried out for different types of finite time controllability. Here we can investigate the case of nonstrict decay, a feature which is not present when considering exponentially decaying functions β. To this end, consider the function $\beta(r, n) = c_n r$ with

$$\begin{array}{lll} \text{(a)} & c_0 = 6, & c_n = 0, \;\; n \geq 1 \\ \text{(b)} & c_0 = c_1 = c_2 = 2, & c_n = 0, \;\; n \geq 3 \\ \text{(c)} & c_0 = c_1 = c_2 = c_3 = 3/2, & c_n = 0, \;\; n \geq 4 \\ \text{(d)} & c_0 = c_1 = c_2 = c_3 = c_4 = c_5 = c_6 = 1, & c_n = 0, \;\; n \geq 7, \end{array} \tag{6.5}$$

which again satisfy $\sum_{n=0}^{\infty} c_n = 6$ and which are depicted in Figure 6.4 for $r = 1$.

Here the respective minimal stabilizing horizons computed from Problem 4.4 evaluate to (a) $\widehat{N} = 11$, (b) $\widehat{N} = 11$, (c) $\widehat{N} = 10$, and (d) $\widehat{N} = 7$. These results confirm the conclusion drawn for the exponentially decaying functions (6.4) (a)–(d), i.e., that fast controllability with large overshoot requires a longer optimization horizon N than slower controllability with smaller overshoot. However, here the differences are less pronounced than in the exponentially decaying case. In fact, the results show that besides the overshoot a decisive feature determining the length of the stabilizing horizon N is the minimal time n_c for which $\beta(r, n_c) < r$, i.e., contraction, can be observed. The longer horizon observed in (6.5)(c) compared to (6.4)(d) is mainly due to the fact that in the former we have $n_c = 1$ while in the latter we have $n_c = 6$.

Finally, we investigate the effect of the weight ω introduced in Remark 4.3. To this end for all the functions from (6.4) and (6.5) we have determined a weight ω such that the corresponding stabilizing optimization horizon $\widehat{N}$ becomes as small as possible. Table 6.1 summarizes our numerical findings.

These results show that suitable tuning of ω reduces the optimization horizon in all cases except for (6.5)(d) (in (6.5)(d), a further reduction to $\widehat{N} < 7$ is not possible because $N = 7$ is the smallest horizon for which controllability to 0 is "visible" in the finite horizon functional J_N). It should, however, be noted that terminal weights

Table 6.1
Minimal stabilizing optimization horizons $\widehat{N}$ for $\omega = 1$ and $\omega > 1$.

Function	$\widehat{N}$ with $\omega = 1$	$\widehat{N}$ with $\omega > 1$	corresponding ω
(6.4)(a)	11	9	9
(6.4)(b)	10	9	5
(6.4)(c)	7	6	3
(6.4)(d)	4	2	6
(6.5)(a)	11	2	6
(6.5)(b)	11	10	4
(6.5)(c)	10	8	25
(6.5)(d)	7	7	arbitrary ≥ 1

$\omega > 1$ have to be used with care, since a wrong choice of ω may also have a destabilizing effect: for instance, using $\omega = 25$ in Case (6.4)(c) leads to $\widehat{N} = 9$ instead of $\widehat{N} = 7$ for $\omega = 1$.

The results also show that (6.3) is no longer the worst case for $\omega > 1$. On the contrary, in the case (6.5)(a) (which is exactly (6.3) for $\gamma = 6$) we obtain the largest reduction of $\widehat{N}$ from 11 to 2.

A reduction to $\widehat{N} = 2$, i.e., to the shortest possible horizon given that $N = 1$ results in a trivial optimal control problem, is possible in cases (6.4)(d) and (6.5)(a). The reason for this is that these two cases exhibit $\beta(r, 1) < r$; i.e., we observe contraction already after one time step. Numerical evidence indicates that stabilization with $N = 2$ and $m = 1$ is always possible in this case. This result actually carries over to the general case $\beta(r, n) < r$ for all $n \geq n_c$ and some $n_c \geq 1$, but only if we increase the control horizon m appropriately: our numerical investigations suggest that in this case we always obtain a stabilizing MPC controller when we chose $N = n_c + 1$, $m = n_c$ and ω sufficiently large, e.g., in example (6.4)(b), where we have $n_c = 2$ we obtain $\widehat{N} = 3$ for $m = 2$ and $\omega = 15$.

In the case just discussed we have $N = m + 1$, i.e., summation up to $N - 1 = m$ in J_N from (2.6), and, thus, the effective optimization horizon coincides with the control horizon. In the PDE optimal control literature, this particular choice of N and m in an MPC scheme is often termed “instantaneous control” (cf., e.g., [9, 10, 12, 18] and the references therein) and, thus, an interesting spin off from our analysis is an additional systems theoretic insight into why and when instantaneous control renders a stable closed loop system.

7. Design of MPC schemes. Our numerical findings from the previous section immediately lead to design guidelines[1] for the choice of l, ω, and m for obtaining stable MPC schemes with small optimization horizons N. These can be summarized as follows:

- design l in such a way that the overshoot $\gamma = \sum_{n=0}^{\infty} \beta(r, n)/r$ becomes as small as possible
- in case of exponential controllability $\beta(r, n) = C\sigma^n r$, reducing the overshoot by reducing C is more efficient than by reducing σ
- in case of finite time controllability $\beta(r, n) = c_n r$, reducing the overshoot by reducing the c_n is more efficient than by reducing the time to reach $l^*(x) = 0$
- terminal weights $\omega > 1$ often lead to smaller N, but too large ω may have the opposite effect, so ω should be tuned with care

[1] These guidelines are derived from numerical evidence by solving Problem 4.4 for a couple of test examples; however, it seems likely that rigorously provable versions could be formulated for most of these statements.

- enlarging m typically leads to smaller N but may decrease the robustness of the closed loop since the feedback is evaluated less frequently
- systems which are contracting after some time n_c, i.e., $\beta(r,n) < r$ for all $n \geq n_c$ are always stabilized by chosing the "instantaneous control" parameters $N = n_c + 1$, $m = n_c$, and ω suffiently large

We illustrate the effectiveness of these guidelines by two examples. We start with a two-dimensional example from [24] given by

$$x(n+1) = \begin{pmatrix} 1 & 1.1 \\ -1.1 & 1 \end{pmatrix} x(n) + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u(n)$$

with running cost

$$l(x,u) = \max\{\|x\|_\infty, |u|\} = \max\{|x_1|, |x_2|, |u|\}.$$

Since this example is low dimensional and linear, V_N can be computed numerically. This fact was used in [24] in order to compute the minimal optimization horizon for a stabilizing MPC feedback law with $m = 1$, which turns out to be $N = 5$ (note that the numbering in [24] differs from ours).

In order to apply our approach, we construct β and u_x meeting Assumption 3.1. Because the system is finite time controllable to 0 this is quite easy to accomplish: using the control

$$u_x(0) = \frac{21}{110}x_1 - 2x_2, \;\; u_x(1) = \frac{221}{110}x_1 + \frac{221}{100}x_2, \;\; u_x(n) = 0, \, n \geq 2$$

for $x(0) = (x_1, x_2)^T$ one obtains the trajectory

$$x_{u_x}(1) = \begin{pmatrix} x_1 + 1.1x_2 \\ -\frac{10}{11}x_1 - x_2 \end{pmatrix}, \;\; x_{u_x}(n) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \, n \geq 2.$$

Since $l^*(x) = \|x\|_\infty$ we can estimate

(7.1)

$$\|x_{u_x}(0)\|_\infty = l^*(x), \;\; \|x_{u_x}(1)\|_\infty \leq 2.1 l^*(x), \;\; |u_x(0)| \leq 2.2 l^*(x), \;\; |u_x(1)| \leq 4.22 l^*(x),$$

implying $l(x_{u_x}(0), u_x(0)) \leq 2.2 l^*(x)$, $l(x_{u_x}(1), u_x(1)) \leq 4.22 l^*(x)$, and $l(x_{u_x}(n), u_x(n)) = 0$ for $n \geq 2$ and, thus, Assumption 3.1 with

$$\beta(r,0) = 2.2\,r, \;\; \beta(r,1) = 4.22\,r, \;\; \beta(r,n) = 0, \, n \geq 2.$$

Solving Problem 4.4 for this β we obtain a minimal stabilizing horizon $N = 12$, which is clearly conservative compared to the value $N = 5$ computed in [24]. Note, however, that instead of using the full information about the functions V_N, which are in general difficult to compute, we only use controllability information on the system.

Now we demonstrate that despite this conservatism our design guidelines can be used to derive a modified design of the MPC scheme which yields stability for horizons $N < 5$. Recall that the estimate for N becomes better, the smaller the overshoot γ is. A look at (7.1) reveals that in this example a reduction of the overshoot can be achieved by reducing the weight of u in l. For instance, if we modify l to

$$l(x,u) = \max\{\|x\|_\infty, |u|/2\},$$

then (7.1) leads to

$$\beta(r,0) = 1.1\,r, \;\; \beta(r,1) = 2.11\,r, \;\; \beta(r,n) = 0, \, n \geq 2.$$

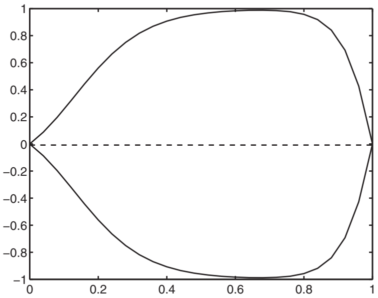


FIG. 7.1. *Equilibria for $u \equiv 0$; solid = asymptotically stable, dashed = unstable.*

Solving Problem 4.4 for this β leads to a minimal stabilizing horizon $N = 5$. Using the terminal weight $\omega = 4$ yields a further reduction to $N = 4$ and if, in addition, we are willing to implement a two-step feedback, i.e., use $m = 2$, then we can reduce the stabilizing optimization horizon even further to $N = 3$. This illustrates how, just by using the controllability information of the system, our analysis can be used to design an MPC scheme reducing the optimization horizon N by 40%.

Our second example demonstrates that our design guidelines are also applicable to infinite dimensional systems. Even though in this case an explicit construction of the controllability function β and the control u_x in Assumption 3.1 is in general rather difficult, we can still apply our results by using the structure of the system equation in order to extract the necessary information about β. To this end, consider the infinite dimensional control system governed by the parabolic reaction-advection-diffusion PDE with distributed control

$$y_t = y_x + \nu y_{xx} + \mu y(y+1)(1-y) + u \tag{7.2}$$

with solutions $y = y(t,x)^2$ for $x \in \Omega = (0,1)$, boundary conditions $y(t,0) = y(t,1) = 0$, initial condition $y(0,x) = y_0(x)$, and distributed control $u(t,\cdot) \in L^2(\Omega)$. The corresponding discrete time system (2.1), whose solutions and control functions we denote by $y(n,x)$ and $u(n,x)$, respectively, is the sampled-data system obtained according to (2.2) with sampling period $T = 0.025$.

For the subsequent numerical computations we discretized the equation in space by finite differences on a grid with nodes $x_i = i/M$, $i = 0, \dots, M$, using backward (i.e., upwind) differences for the advection part y_x. Figure 7.1 shows the equilibria of the discretized system for $u \equiv 0$, $\nu = 0.1$, $\mu = 10$, and $M = 25$.

Our goal is to stabilize the unstable equilibrium $y^* \equiv 0$, which is possible because with the additive distributed control we can compensate the whole dynamics of the system. In order to achieve this task, a natural choice for a running cost l is the tracking type functional

$$l(y(n,\cdot), u(n,\cdot)) = \|y(n,\cdot)\|^2_{L^2(\Omega)} + \lambda \|u(n,\cdot)\|^2_{L^2(\Omega)}, \tag{7.3}$$

which we implemented with $\lambda = 10^{-1}$ for the discretized model in MATLAB using the `lsqnonlin` solver for the resulting optimization problem.

[2] Note the change in the notation: x is the independent state variable while $y(t,\cdot)$ is the new state, i.e., X is now an infinite dimensional space.

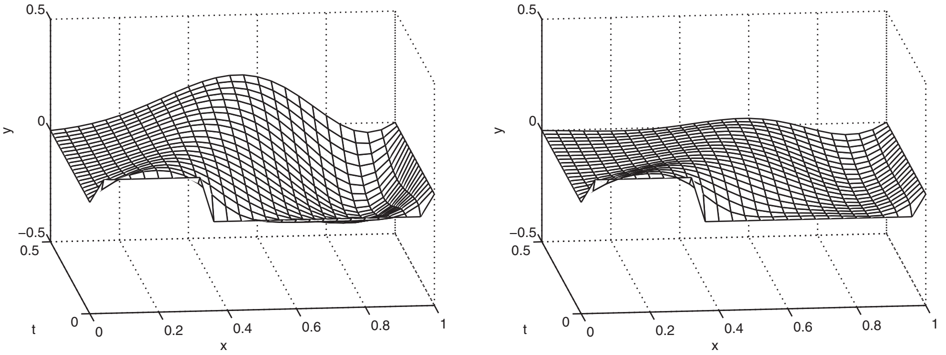


FIG. 7.2. *Receding horizon with l from* (7.3)*,* $N = 3$ *(left) and* $N = 11$ *(right).*

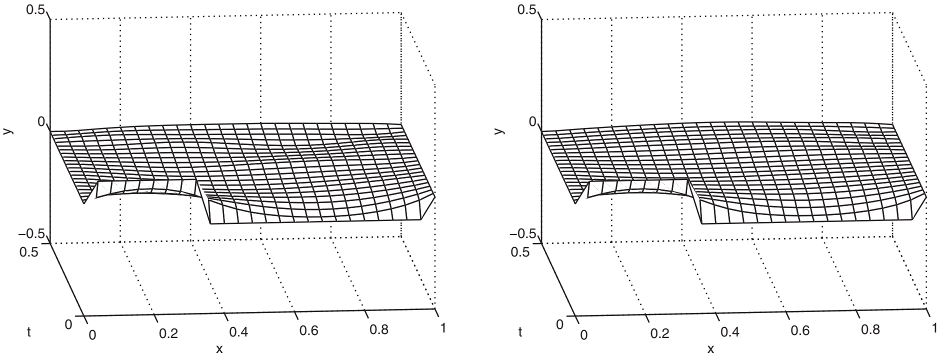


FIG. 7.3. *Receding horizon with l from* (7.4)*,* $N = 2$ *(left) and* $N = 3$ *(right).*

The simulations shown in Figure 7.2 reveal that the performance of this controller is not completely satisfactory: for $N = 11$ the solution remains close to $y^* = 0$ but does not converge while for $N = 3$ the solution even grows.

The reason for this behavior lies in the fact that in order to control the system to $y^* = 0$, in (7.2) the control needs to compensate for y_x, i.e., any stabilizing control must satisfy $\|u(n,\cdot)\|^2_{L^2(\Omega)} \gtrsim \|y_x(n,\cdot)\|^2_{L^2(\Omega)}$. Thus, for any stabilizing control sequence u we obtain $J_\infty(y_0, u) \gtrsim \lambda\|y_x(n,\cdot)\|^2_{L^2(\Omega)}$ which—even for small values of λ—may be considerably larger than $l^*(y) = \|y\|^2_{L^2(\Omega)}$, resulting in a large β and, thus, the need for a large optimization horizon N in order to achieve stability.

This effect can be avoided by changing l in such a way that $l^*(y)$ includes $\|y_x\|^2_{L^2(\Omega)}$, e.g., by setting

$$(7.4) \qquad l(y(n,\cdot), u(n,\cdot)) = \|y(n,\cdot)\|^2_{L^2(\Omega)} + \|y_x(n,\cdot)\|^2_{L^2(\Omega)} + \lambda\|u(n,\cdot)\|^2_{L^2(\Omega)}.$$

For this l the control effort needed in order to control (7.2) to $y^* = 0$ is proportional to $l^*(y)$. Thus, the overshoot reflected in the controllability function β is now essentially proportional to $1 + \lambda$ and, thus, in particular, small for our choice of $\lambda = 10^{-1}$ which implies stability even for small optimization horizon N. The simulations for the corresponding discretized running cost are illustrated in Figure 7.3 and show that this is, indeed, the case: we obtain asymptotic stability even for the very small optimization horizons $N = 2$ (i.e., for instantaneous control) and $N = 3$, with slightly better performance for the latter case.

8. Conclusions and outlook. We have presented a stability and performance analysis technique for unconstrained nonlinear MPC schemes which relies on a suitable controllability condition for the running cost. The proposed technique leads to a small optimization problem whose size depends only on the optimization horizon to be investigated but not on the dimension of the state space. The stability condition obtained this way turns out to be tight with respect to the class of systems satisfying the assumed controllability condition. The numerical analysis based on this optimization problem was used to derive guidelines for the design of MPC schemes guaranteeing stability for small optimization horizons N. The effectiveness of these guidelines has been illustrated by a finite and an infinite dimensional example.

Future research will include the generalization of the approach to situations where V_N cannot be expected to be a Lyapunov function, the inclusion of deterministic and stochastic uncertainties in the analysis and the relaxation of the Assumptions 3.1 and 5.1(ii) to more general controllability and detectability assumptions.

9. Appendix: A technical lemma.

LEMMA 9.1. *Consider $\beta \in \mathcal{KL}_0$, $N \geq 1$, $m \in \{1, \dots, N-1\}$, a sequence $\lambda_n > 0$, $n = 0, \dots, N-1$ and $\nu > 0$ fulfilling* (4.1), (4.2) *and*

$$\sum_{n=0}^{N-1} \lambda_n - \nu \leq \alpha \sum_{n=0}^{m-1} \lambda_n \tag{9.1}$$

for some $\alpha < 0$. Then there exist $\bar{\lambda}_n > 0$, $\bar{\nu} > 0$, and $\bar{\alpha} < 0$ satisfying (4.1), (4.2), *and* (9.1) *for which the inequalities* (4.1) *are strict.*

Proof. We label the inequalities for $\bar{\lambda}_n$, $\bar{\nu}$, and $\bar{\alpha}$ by $(\overline{4.1})$, $(\overline{4.2})$, and $(\overline{9.1})$, respectively, and set $\bar{\lambda}_n = \lambda_n$, $n = 0, \dots, N-2$ and $\bar{\lambda}_{N-1} = \lambda_{N-1} - \varepsilon$ where $\varepsilon \in (0, \lambda_{N-1})$ is specified below. Since this implies $\bar{\lambda}_{N-1} < \lambda_{N-1}$ the inequalities $(\overline{4.1})$ are strict. Furthermore, $(\overline{9.1})$ holds for all $\bar{\alpha} \geq \alpha$ and $(\overline{4.2})$ holds for $j = 1, \dots, N-m-2$.

It, thus, remains to choose ε, $\bar{\nu}$, and $\bar{\alpha}$ such that $(\overline{4.2})$ holds for $j = N-m-1$ while $(\overline{9.1})$ and $(\overline{4.2})$ for $j = 1, \dots, N-m-2$ remain valid. In case the inequality (4.2) for $j = N-m-1$ is strict, we choose $\bar{\nu} = \nu$, $\bar{\alpha} = \alpha$, and $\varepsilon > 0$ sufficiently small such that $(\overline{4.2})$ holds for $j = N-m-1$, which is possible since B_k is continuous.

In case that (4.2) for $j = N-m-1$ is an equality, we set $\bar{\nu}$ (depending on ε) such that equality in $(\overline{4.2})$ for $j = N-m-1$ holds, as well. This implies $\bar{\nu} \leq \nu$ and, thus, all other inequalities in $(\overline{4.2})$ remain valid for all $\varepsilon \in (0, \lambda_{N-1})$. Now by continuity of B_k the value $\bar{\nu}$ depends continuously on ε; hence, for $\varepsilon > 0$ sufficiently small we obtain $(\overline{9.1})$ for $\bar{\alpha} = \alpha/2 < 0$. □

Acknowledgment. I would like to thank Karl Worthmann for his help with the numerical computations for the PDE example (7.2).

REFERENCES

[1] F. ALLGÖWER AND A. ZHENG, EDS., *Nonlinear model predictive control*, Progress in Systems and Control Theory 26, Birkhäuser Verlag, Basel, 2000.

[2] E. F. CAMACHO AND C. BORDONS, *Model Predictive Control*, 2nd ed., Springer-Verlag, London, 2004.

[3] H. CHEN AND F. ALLGÖWER, *A quasi-infinite horizon nonlinear model predictive control scheme with guaranteed stability*, Automatica, 34 (1998), pp. 1205–1217.

[4] G. GRIMM, M. J. MESSINA, S. E. TUNA, AND A. R. TEEL, *Model predictive control: For want of a local control Lyapunov function, all is not lost*, IEEE Trans. Automat. Control, 50 (2005), pp. 546–558.

[5] L. Grüne, *Computing stability and performance bounds for unconstrained NMPC schemes*, in Proceedings of the 46th IEEE Conference on Decision and Control, New Orleans, LA, 2007, pp. 1263–1268.

[6] L. Grüne and J. Pannek, *Practical NMPC suboptimality estimates along trajectories*, Systems Control Lett., 58 (2009), pp. 161–168.

[7] L. Grüne and A. Rantzer, *Suboptimality estimates for receding horizon controllers*, in Proceedings of the 17th International Symposium on Mathematical Theory of Networks and Systems MTNS2006, Kyoto, Japan, 2006, pp. 120–127.

[8] L. Grüne and A. Rantzer, *On the infinite horizon performance of receding horizon controllers*, IEEE Trans. Automat. Control, 53 (2008), pp. 2100–2111.

[9] M. Hinze, *Instantaneous closed loop control of the Navier-Stokes system*, SIAM J. Control Optim., 44 (2005), pp. 564–583.

[10] M. Hinze and S. Volkwein, *Analysis of instantaneous control for the Burgers equation*, Nonlinear Anal., 50 (2002), pp. 1–26.

[11] B. Hu and A. Linnemann, *Toward infinite-horizon optimality in nonlinear model predictive control*, IEEE Trans. Automat. Control, 47 (2002), pp. 679–682.

[12] R. Hundhammer and G. Leugering, *Instantaneous control of vibrating string networks*, in Online optimization of large scale systems, M. Grötschel, S. O. Krumke, and J. Rambau, eds., Springer, Berlin, 2001, pp. 229–249.

[13] A. Jadbabaie and J. Hauser, *On the stability of receding horizon control with a general terminal cost*, IEEE Trans. Automat. Control, 50 (2005), pp. 674–678.

[14] S. S. Keerthy and E. G. Gilbert, *Optimal infinite horizon feedback laws for a general class of constrained discrete-time systems: Stability and moving horizon approximations*, J. Optimiz. Theory Appl., 57 (1988), pp. 265–293.

[15] H. K. Khalil, *Nonlinear Systems*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1996.

[16] B. Lincoln and A. Rantzer, *Relaxing dynamic programming*, IEEE Trans. Autom. Control, 51 (2006), pp. 1249–1260.

[17] L. Magni, G. De Nicolao, L. Magnani, and R. Scattolini, *A stabilizing model-based predictive control algorithm for nonlinear systems*, Automatica, 37 (2001), pp. 1351–1362.

[18] X. Marduel and K. Kunisch, *Suboptimal control of transient nonisothermal viscoelastic fluid flows*, Phys. Fluids, 13 (2001), pp. 2478–2491.

[19] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. M. Scokaert, *Constrained model predictive control: Stability and optimality*, Automatica, 36 (2000), pp. 789–814.

[20] D. Nešić and A. R. Teel, *A framework for stabilization of nonlinear sampled-data systems based on their approximate discrete-time models*, IEEE Trans. Automat. Control, 49 (2004), pp. 1103–1122.

[21] J. A. Primbs and V. Nevistić, *Feasibility and stability of constrained finite receding horizon control*, Automatica, 36 (2000), pp. 965–971.

[22] S. J. Qin and T. A. Badgwell, *An overview of nonlinear model predictive control applications*, in Nonlinear model predictive control, J. C. Cantor, C. E. Garcia, and B. Carnahan, eds., Birkhäuser, Basel, 2000.

[23] A. Rantzer, *Relaxed dynamic programming in switching systems*, IEEE Proceedings — Control Theory and Applications, 153 (2006), pp. 567–574.

[24] J. S. Shamma and D. Xiong, *Linear nonquadratic optimal control*, IEEE Trans. Automat. Control, 42 (1997), pp. 875–879.

[25] E. D. Sontag, *Comments on integral variants of ISS*, Syst. Control Lett., 34 (1998), pp. 93–100.